NAVAL RESEARCH LOGISTICS QUARTERLY

JUNE 1980
VOL. 27, NO. 2

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $11.15 a year in the U.S. and Canada, $13.95 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, P-35 (Revised 1-74).

A SINGLE PERIOD MODEL FOR A MULTIPRODUCT PERISHABLE INVENTORY SYSTEM WITH ECONOMIC SUBSTITUTION

Bryan L. Deuermeyer

Department of Industrial Engineering
Texas A&M University
College Station, Texas 77843

ABSTRACT

This paper develops a single period model for a specific class of multiproduct perishable inventory systems where demands are interdependent. This class of inventory systems has the property that there is economic substitution between products. It is shown that the optimal policy has the economic substitution property, and that the rate of substitution is age dependent. The model serves as a generalization of a theorem discovered by Ignall and Veinott.

1. INTRODUCTION

This paper considers the problem of characterizing the properties of optimal ordering policies for a specific class of single period, multiproduct, perishable inventory systems where there may be dependencies between the stochastic product demands. It is assumed that there are n products and m demand classes for these products and that each product has a known fixed (and finite) lifetime. As is usual, the criterion for choosing an optimal ordering policy is the minimization of total expected costs. Pertinent costs are: linear purchasing (production) and outdating (disposal) costs and convex holding and shortage costs.

The special class we consider is determined by the property called economic substitution. By this, it is meant that the order quantity of product $i (i = 1, 2, \ldots, n)$ is a nonincreasing function of the on-hand inventories of the other products. In effect, some amount of product i can be replaced by an increase in the amounts of other products kept on hand, in terms of the economic benefits realized by the firm. This phenomenon typically occurs because space and capital allocated to, say, product i must be adjusted (downward) when increasing the inventory commitment to another product. It is important to emphasize that the phrase "economic substitution" as it is used here does not imply that another product, say j, can be substituted for product i to satisfy a demand for product i. (The latter would correspond to the classical economic interpretation.)

The model developed here would find applications in the retail food industry, photographic film industry, pharmaceutical industry, and finally in hospital and regional blood banks.

Our focus is to characterize the properties of the optimal policy when economic substitution is in effect. We show in Section 3 that the optimal starting inventory (on-hand plus amount ordered) is in general more responsive to changes in newer inventory than older

inventory. This phenomenon carries over to economic substitution. That is, the rate of substitution is age dependent.

The study of perishable inventory theory has been an active research area in recent years. Van Zyl [11] provided the first major study on single-product perishable inventories, but the field only began to develop after the work of Fries [5] and Nahmias [7]. These three papers all dealt with the determination of optimal policies. These policies were very complicated to implement, particularly from a computational standpoint. This motivated Cohen [2] to develop an approach for determining optimal single critical number policies that were easy to use. Chazan and Gal [1] proved a conjecture made in [2] that the expected number of outdates is a convex function. More recently, Nahmias [9] developed an approach that allows relatively easy computation of myopic approximate policies.

Very little work has been done in the area of multiproduct perishable inventory theory. Nahmias and Pierskalla [10] considered a two-product perishable/nonperishable model which is to some extent applicable to whole blood/frozen blood inventories, and to foodstuffs such as milk and dry milk. Deuermeyer [3] considered a two-product perishable inventory-production model where different production processes must be coordinated to make the inventory items. The objective of this paper is to characterize the age-dependent relationships between products.

Several papers have been written in the area of multiproduct (nonperishable) inventory theory. A summary of early research can be found in the excellent survey by Veinott [12]. Deuermeyer and Pierskalla [4] considered a two-product inventory-production system where a by-product process and a single-item production process must be coordinated to manufacture the two products. Ignall and Veinott [6] originated the concept of economic substitution that we use here.

The article is organized as follows: Section 2 provides the notation and assumptions upon which the model is based; Section 3 provides a detailed discussion of the optimal policy for the general one-period model, and the special case for $n = 2$ is presented as an example.

2. NOTATION AND ASSUMPTIONS

Let the number of products be n and the number of demand classes be m. Then the vector random variable $D = (D_1, D_2, \ldots, D_m)$ is the joint demand during the period. We assume that the range of D is a Borel set $\mathbb{D}$. From D we form the product demand vector $\bar{D} = (\bar{D}_1, \ldots, \bar{D}_n)$ where $\bar{D}_i$ is the total demand for product i during the period. Let l_i be the fixed lifetime for product i, $i \in I = \{1, 2, \ldots, n\}$. Let $l = l_1 + l_2 + \ldots + l_n - n$. l is the dimension of the statespace of the problem.

The following assumptions and additional notation are required:

1. l_i is integral valued and $1 < l_i < \infty$, $i \in I$.
2. All stock is issued to meet demand according to a FIFO policy (that is, oldest first).
3. $\bar{D}$ has a continuous density of f and distribution F such that $f(t)$ is zero except on the positive orthant of E_n (Euclidean n-space).
4. All units are fresh when they enter stock and there is no delivery lead time.
5. The vector of total starting (on-hand plus amount ordered and before demand) inventories z must lie in a topologically closed set Z.
6. Let $L(z)$, $z \in Z$ be the expected inventory holding and shortage cost function. Specific assumptions concerning $L(\cdot)$ will be made subsequently.

7. Let c_i be the proportional cost per unit of product i purchased, $i \in I$.

8. Let θ_i be the proportional cost charged against each unit of product i that outdates, $i \in I$. The exact method for charging these costs will be clarified shortly.

Let x_{ij} be the initial inventory (already on hand at the beginning of the period) of age j, $j \in [i] = \{1, 2, \ldots, l_i - 1\}$ of product i, $i \in I$. Further, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{il_i-1})$ be the vector of initial inventories of products i by age, $i \in I$. The notation $\mathbf{x}$ will be reserved for the vector $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ of inventories by age. In addition, we define

$$x_i = \sum_{j \in [i]} x_{ij}, \quad i \in I,$$

which is the total amount of product i on hand prior to ordering. Finally let $x = (x_1, \ldots, x_n)$. Once again $\mathbf{x}$ refers to inventory by age while x simply refers to cumulative stock.

Let $y_i \geqslant 0$ be the amount of product i ordered, $i \in I$, and $y = (y_1, \ldots, y_n)$ be the vector of order quantities. Then, the vector of starting inventories (after ordering but before demand) is given by $z = x + y$. We require that $z \in Z$.

Let a and b be vectors in E_n. $a \leqq b$ means that $a_i \leqq b_i$, $i \in I$, and $a \leqslant b$ means that $a \leqq b$ but $a_i < b_i$ for some i. Similarly, $a < b$ means that $a_i < b_i$, $i \in I$. Let $[a,b] = \{x \in E_n;\ a_i \leqq x_i \leqq b_i,\ i \in I\}$. $[a,b]$ is called a closed rectangle. Finally, let E_n^* be the extended Euclidean space. If H is a matrix, then $|H|$ is the determinant of H. Let H be a square matrix. H_{ik} is the matrix formed from H by deleting row i and column k and H^{ik} is the matrix formed from H by interchanging columns i and k.

Let ϕ be a function defined on E_n that is twice continuously differentiable over E_n. Then we write

$$D_i \phi\ (x_1, \ldots, x_n) = \frac{\partial}{\partial x_i} \phi\ (x_1, \ldots, x_n), \quad i = 1, 2, \ldots, n,$$

and

$$D_{ij} \phi\ (x_1, \ldots, x_n) = \frac{\partial}{\partial x_j} D_i \phi\ (x_1, \ldots, x_n), \quad i, j = 1, 2, \ldots, n.$$

We let $\nabla^2 \phi\ (x)$ be the Hessian matrix of $\phi(\cdot)$ evaluated at x. Also, let t be fixed and let $y = (y_1, \ldots, y_t)$ and $x = (y_{t+1}, \ldots, y_n)$. Then, $\nabla_y^2 \phi\ (y,x)$ is the Hessian of ϕ restricted to the first t variables.

We follow the method of Nahmias [7] for assigning the outdate cost; that is, the outdate cost θ_i is charged against the expected amount of the order y_i that outdates l_i periods into the future, given $\mathbf{x}_i$ on hand. Then the appropriate cost function is given by

$$V_i\ (y_i, \mathbf{x}_i) = \theta_i \int_0^{y_i} G_i\ (u, \mathbf{x}_i)\, du,$$

where $G_i\ (\cdot, \mathbf{x}_i)$ is computed recursively using the marginal distribution $F_i(\cdot)$ of the total demand for product i. For the complete derivation of the "V" functions and their properties, see [7], and for computational results when demands are Erlang, see [8].

The motivation for this approach is that the full impact of outdating is taken into account at the time the order is placed. This approach is particularly attractive when studying single-period inventory models as we do in Section 3, since the one-period optimal solution would otherwise ignore outdating entirely.

An important concept used extensively throughout this paper is that of a substitute matrix, first defined by Ignall and Veinott [6]. Let $H(nxn)$ be a nonnegative symmetric positive definite matrix. Let H^{ij} be the matrix formed from H by interchanging columns i and j, where $i < j$. Then we say that H is a substitute matrix if every principal minor of H^{ij} that contains elements from only one of the columns i or j is nonnegative. We let S_n be the class of all substitute matrices of order n. When $n = 2$, S_n consists of all positive definite symmetric matrices with nonnegative off-diagonal elements. Unfortunately, S_n is closed under addition only when $n = 2$ (see [6] for examples).

3. ANALYSIS OF OPTIMAL POLICIES

The primary aim of this article is to show that certain assumptions will be sufficient for the optimal policy to have the economic substitution property. In addition, we will demonstrate the additional properties of the optimal policy that are obtained as a consequence of perishability.

We define the term optimal policy to be the specific choice of starting inventory level that leads to the minimum expected total cost. If such a policy exists, we denote it as $z(\mathbf{x}) = (z_1(\mathbf{x}),\ldots,z_n(\mathbf{x}))$. Thus, $z(\mathbf{x})$ solves

$$B(z(\mathbf{x});\mathbf{x}) = \inf_{z \in Z} B(z;\mathbf{x}), \quad \mathbf{x} \in E_l \tag{3.1}$$

with

$$B(z;\mathbf{x}) = G(z) + \sum_{i \in I} V_i(z_i - x_i;\mathbf{x})\dagger$$

and

$$G(z) = \sum_{i \in I} c_i z_i + L(z), \quad z \in Z.$$

We propose the following two postulates:

(A1) Assume that $\nabla_z B(z;\mathbf{x})$ is a substitute matrix for each $\mathbf{x} \in E_l$.

(A2) Assume that all sets $A(s) = \{z \in Z;\ G(z) \leqq s\}$ are compact for each bounded real number s.

The assumption that $\nabla_z^2 B(z;\mathbf{x})$ be a substitute matrix is only slightly more restrictive than requiring that $B(\cdot;\mathbf{x})$ be strictly convex with nonnegative crosspartial derivatives. The latter occurs whenever the marginal total costs of product i are nondecreasing functions of the inventories of all other products. This assumption turns out to be sufficient for the optimal ordering policy to have the economic substitution property. Postulate (A2) will hold under a number of conditions provided postulate (A1) is satisfied. If, for example, one of the sets $A(s)$ is bounded for some s', it is a property of convex functions that $A(s)$ will be bounded for every s. A sufficient condition is to require that $G(y) \to \infty$ whenever $||y|| \to \infty$, where $||\cdot||$ is the Euclidean norm on E_n. Finally, the compactness will follow by the boundedness of $A(\cdot)$ and the continuity of $G(\cdot)$. Postulate (A2) assures that the finite optimal policy exists and that "inf" can be replaced by "min" in Eq. (3.1).

†Notice that in our definition of $B(z;\mathbf{x})$ we have not included the constant $-\sum c_i x_i$. This is justified since the constant does not affect the optimization.

The central result of this article is Theorem 3.1. It generalizes Theorem 6 in Ignall and Veinott [6] to perishable inventory products.

THEOREM 3.1 If postulates (A1) and (A2) hold and $Z = [a,b] \subseteq E_n^*$, a unique $z(\mathbf{x})$ exists and is continuously differentiable. In addition, suppose $z_i(\mathbf{x}) > x_i$ for $i \in I' \subseteq I$, and $z_i(\mathbf{x}) = x_i$ otherwise. Let $i \in I'$, $k \in I$, and $m \in [k] - \{2\}$. Then, the optimal policy has the following properties:

$$\text{(P1)} \qquad \frac{\partial z_i}{\partial x_k} > 0, \quad k = i,$$

$$\text{(P2)} \qquad \frac{\partial z_i}{\partial x_k} \leqq 0, \quad k \neq i,$$

$$\text{(P3)} \qquad \frac{\partial z_i}{\partial x_{km-1}} - \frac{\partial z_i}{\partial x_{km}} \leqq 0, \quad k = i,$$

$$\text{(P4)} \qquad \frac{\partial z_i}{\partial x_{km-1}} - \frac{\partial z_i}{\partial x_{km}} \geqq 0, \quad k \in I', \; k \neq i,$$

$$\text{(P5)} \qquad \frac{\partial z_i}{\partial x_{km-1}} - \frac{\partial z_i}{\partial x_{km}} = 0, \quad k \notin I',$$

where $z_i \equiv z_i(\mathbf{x})$.

Before proving Theorem 3.1, it is important to interpret the five properties of the optimal policy. Property (P2) shows that the optimal policy has the economic substitution property; the optimal starting inventory of product i is a nonincreasing function of the initial inventory of product k, $k \neq i$. This is sufficient to prove that the ordering quantity for product i is nonincreasing. This result is the analog of the assertion of Theorem 6 in Ignall and Veinott [6]. Properties (P1) and (P3)-(P5) provide the additional properties that arise from perishability.

Properties (P1) and (P3) provide the extension of earlier work in perishable inventory theory on single-product models to multiproduct situations. Property (P1) shows that the optimal starting inventory of a product is a nondecreasing function of its initial inventory. That is, the ordering quantity strictly decreases when initial inventory increases, but not enough to offset the increase. Property (P3) shows that the optimal starting inventory of a product is, in general, more sensitive to changes in younger units than in older units. These two results were discovered by Fries [5] and Nahmias [7] in the single-product case.

Properties (P4) and (P5) demonstrate that the economic substitution property is also affected by making changes in the age distribution of the units in stock. Property (P4) shows that the rate of substitution of product k for product i is more sensitive to increases in the younger units of product k than in older units. However, Property (P5) shows that this age-dependent rate of substitution only holds for the products that have been ordered—not all products.

PROOF: The logic required to prove Theorem 3.1 is essentially that used by Ignall and Veinott [6], except that we need to account for the fact that the state variable is $\mathbf{x}$, not x, and the objective function depends on the state variable through the outdating cost function. All that is needed to make their proof apply to the present theorem is to generalize their Lemma 7, so that it incorporates the above information. Once this is accomplished, their proof only needs

minor changes. For this reason, we do not describe a proof of Theorem 3.1 here, but instead focus on the proof of the new lemma, Lemma 3.1.

Q.E.D.

Before we can state the lemma, some additional notation is required. Let t be fixed and $1 \leqq t < n$, and let $i, j \in I$. Let $h_{ij} = D_{ij}\, B(z;\mathbf{x})$, $h_i = (h_{i1},\ldots,h_{it})$, and H be the matrix with rows h_i. Also, let H_i be the matrix formed from H by replacing column i by h_k' ($'$ denotes the transpose operation).

Now, let $\bar{x} = (x_{t+1},\ldots,x_n)$ and let $y(\mathbf{x}) = (y_1(\mathbf{x}),y_2(\mathbf{x}),\ldots,y_t(\mathbf{x}))$ solve

$$B(y(\mathbf{x}),\bar{x};\mathbf{x}) = \min_{y} B(y,\bar{x};\mathbf{x}).$$

Also, let

$$u_j(m) = \frac{\partial y(\mathbf{x})}{\partial x_{km}},\ m \in [k], \text{ and } u(m) = (u_1(m),\ldots,u_t(m))'.$$

LEMMA 3.1 (Generalization of Lemma 7 in Ignall and Veinott [6]): Let $y \equiv y(\mathbf{x})$.

1. Let $1 \leqq i \leqq t$, $m \in [i]$. Then,

   (a) $\dfrac{\partial y_i}{\partial x_{im}} > 0,$

   (b) $\dfrac{\partial y_i}{\partial x_{im-1}} - \dfrac{\partial y_i}{\partial x_{im}} \leqq 0,$ for $m \neq 2$.

2. Let $1 \leqq i \leqq t$, $1 \leqq k \leqq n$, $k \neq i$, and $m \in [k]$. Then,

   (a) $\dfrac{\partial y_i}{\partial x_{km}} \leqq 0,$

   (b) $\dfrac{\partial y_i}{\partial x_{km-1}} - \dfrac{\partial y_i}{\partial x_{km}} \geqq 0,\ m \neq 2.$

3. Let $1 \leqq i \leqq t$, $t+1 \leqq k \leqq n$, $m \in [k] - \{2\}$. Then,

$$\frac{\partial y_i}{\partial x_{km-1}} - \frac{\partial y_i}{\partial x_{km}} = 0.$$

4. Let $t+1 \leqq i \leqq n$, $t+1 \leqq k \leqq n$, $m \in [k]$. Then,

$$\frac{\partial}{\partial x_{km}} D_i\, B(y(\mathbf{x}),\bar{x};\mathbf{x}) \geqq 0.$$

The proof of Lemma 3.1 requires a lemma. This we put first.

LEMMA 3.2 Let $1 \leqq i \leqq n$, $1 \leqq k \leqq n$, $m \in [k]$. Then,

$$\frac{\partial}{\partial x_{km}} D_i\, B(z;\mathbf{x}) = \begin{cases} 0, & k \neq i \\ -\Delta_k, & k = i \end{cases}$$

where $\Delta_{km} = D_{1m+1}\, V_k(z_k - x_k;\mathbf{x}_k) - D_{11}\, V_k(z_k - x_k;\mathbf{x}_k)$.

PROOF: We simply state the relevant partial derivatives of B:

$$D_i\, B(z;\mathbf{x}) = D_i\, G(z) + D_1 V_i(z_i - x_i;\mathbf{x}_i),$$

$$D_{ii}\, B(z;\mathbf{x}) = D_{ii}\, G(z) + D_{11} V_i(z_i - x_i;\mathbf{x}_i),$$

$$D_{ik}\, B(z;\mathbf{x}) = D_{ik}\, G(z),\ k \neq i,$$

$$\frac{\partial}{x_{km}} D_i\, B(z;\mathbf{x}) = \begin{cases} 0, & k \neq i \\ -D_{11}\, V_i(z_i - x_i;\mathbf{x}_i) + D_{1m+1} V_i(z_i - x_i;\mathbf{x}_i), & k = i. \end{cases}$$

Q.E.D.

PROOF (LEMMA 3.1): Fix i and k such that $1 \leqq i \leqq t$, $1 \leqq k \leqq t$, and let $m \in [k]$. Also, let e_k be the t-dimensional unit vector with a 1 in position k. Then, using Lemma 3.2 it can be shown that

$$Hu(m) = \Delta_{km}\, e_k$$

so that

$$u_i(m) = \begin{cases} -\Delta_{km} \cdot |H_{kk}^{ik}|/|H|, & i \neq k \\ \Delta_{km}\, |H_{kk}|/|H|, & i = k. \end{cases}$$

Now $|H| > 0$, $|H_{kk}| > 0$, $|H_{kk}^{ik}| \geqq 0$ from postulate (A1) and by properties of substitute matrices. Finally, $\Delta_{km} \geqq 0$ from Nahmias [7]. This establishes 1(a). Now,

$$u_i(m-1) - u_i(m) = \begin{cases} [D_{1m+1}\, V_k - D_{1m}\, V_k] \cdot |H_{kk}^{ik}|/|H|, & i \neq k \\ [D_{1m}\ \ V_k - D_{1m+1}\, V_k] \cdot |H_{kk}|/|H|, & i = k, \end{cases}$$

where $V_k = V_k(z_k(\mathbf{x}) - x_k;\mathbf{x}_k)$.

Therefore,

$$u_i(m-1) - u_i(m) \begin{cases} \geqq 0, & i \neq k \\ \leqq 0, & 1 = k \end{cases}$$

due to results in [7]. This establishes 1(b) and 2(b).

Now, fix i and k such that $1 \leqq i \leqq t$ and $t+1 \leqq k \leqq n$, and let $m \in [k]$. Then, Ignall and Veinott [6] show that

$$u_i(m) = -\,|H_i|/|H| \leqq 0.$$

But, this relation also shows that $u_i(m-1) - u_i(m) = 0$. This establishes 2(a) and 3.

Now, fix i and k such that $t+1 \leqq i,\ k \leqq n$, and let $m \in [k]$. Also, let

$$\mathrm{v}(m) = \frac{\partial}{\partial x_{km}} D_i\, B\ (y(\mathbf{x}), \bar{x};\mathbf{x}).$$

Then, analogously to [6], we obtain that

$$\mathrm{v}(m) = \frac{\begin{vmatrix} H & h_k \\ h_i & h_{ik} \end{vmatrix}}{|H|}.$$

Therefore, $v(m)$ is nonnegative (positive if $k = i$), independent of m. This establishes 4 and completes the proof.

Q.E.D.

An interesting special case corresponds to $n = 2$. In this case, it is possible to graph the decision regions and observe the economic substitution and the impact of perishability on the ordering policy. The optimal policy is defined by four regions R_1, R_2, R_{12}, and R_0. These in turn are defined by a point z^* and two functions $y_1(\mathbf{x}_2)$ and $y_2(\mathbf{x}_1)$. The point $z^* = (z_1^\bullet, z_2^\bullet)$ is the minimum of $G(\cdot)$ over Z. $y_1(\mathbf{x}_2)$ is defined as the solution to

$$G^{(1)}(y_1(\mathbf{x}_2), x_2) = 0 \quad \text{for } x_2 \geqq z_2^\bullet$$

and

$$y_1(\mathbf{x}_2) = z_1^\bullet \quad \text{for } x_2 < z_2^\bullet.$$

Similarly, $y_2(\mathbf{x}_1)$ is defined as the solution to

$$G^{(2)}(x_1, y_2(\mathbf{x}_1)) = 0 \quad \text{for } x_1 \geqq z_1^\bullet$$

and

$$y_2(\mathbf{x}_1) = z_2^\bullet \quad \text{for } x_1 < z_1^\bullet.$$

The following two corollaries to Theorem 3.1 provide additional properties of the optimal policy in this special case.

COROLLARY 3.1: If $z(\mathbf{x}) > x$, then $z(\mathbf{x}) \leqslant z^*$. (This rules out the possibility of $z(\mathbf{x}) = z^*$.)

COROLLARY 3.2: The regions characterizing the optimal policy are defined as follows:

1. $\mathbf{x} \in R_1 = \{\mathbf{x} \in E_l; x_1 < y_1(\mathbf{x}_2), x_2 \geqq z_2^\bullet\}$ iff $z_2(\mathbf{x}) = x_2$ and $x_1 < z_1(\mathbf{x}) < y_1(\mathbf{x}_2)$.
2. $\mathbf{x} \in R_2 = \{\mathbf{x} \in E_l; x_2 < y_2(\mathbf{x}_1), x_1 \geqq z_1^\bullet\}$ iff $z_1(\mathbf{x}) = x_1$ and $x_2 < z_2(\mathbf{x}) < y_2(\mathbf{x}_1)$.
3. $\mathbf{x} \in R_{12} = \{\mathbf{x} \in E_l; x < z^*\}$ iff $x < z(\mathbf{x}) \leqslant z^*$.
4. $\mathbf{x} \in R_0 = (R_1 \cup R_2 \cup R_{12})^c$ iff $z(\mathbf{x}) = x$.

When $l_1 = l_2 = 2$, the results of the two corollaries can be depicted as in Fig. 1. Notice that $z_1(\cdot)$ and $z_2(\cdot)$ will be nonincreasing.

4. CONCLUSIONS

The model presented and studied in this article extends the work by Ignall and Veinott [6] to the perishable product case. The primary result in this paper shows that when the products are perishable, the optimal policy has age-dependent economic substitution, under the assumptions given. As we show in a subsequent work, the policies developed here provide a convenient basis for constructing approximately optimal policies for the dynamic problem.

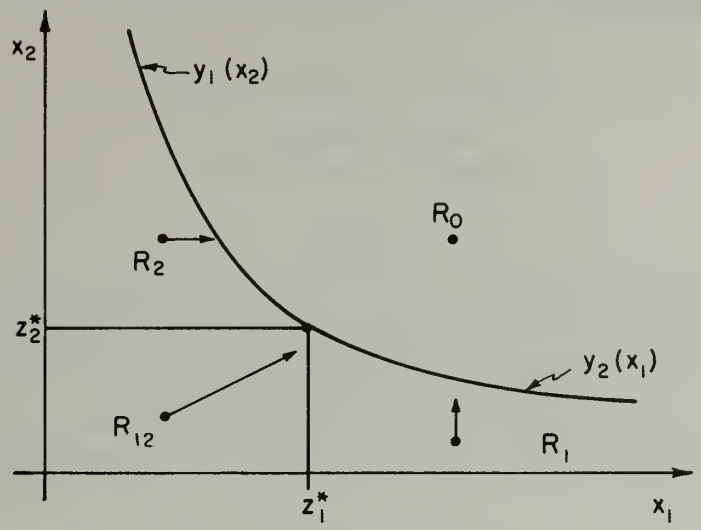

FIGURE 1. Characterization of the optimal policy when $n = l_1 = l_2 = 2$.

REFERENCES

[1] Chazan, D. and S. Gal, "Markovian Model for Perishable Product Inventory," Management Science *23*, 512-521 (1977).

[2] Cohen, M.A., "Analysis of Single Critical Number Ordering Policies for Perishable Inventories," Operations Research *24*, 726-741 (1976).

[3] Deuermeyer, B.L., "A Multi-Type Production System for Perishable Inventories," Operations Research, *27*, 935-943 (1979).

[4] Deuermeyer, B.L. and W.P. Pierskalla, "A By-Product Production System with an Alternative," Management Science *24*, 1173-1183 (1978).

[5] Fries, B.E., "Optimal Ordering Policy for a Perishable Commodity with Fixed Lifetime," Operations Research *23*, 62-74 (1975).

[6] Ignall, E. and A.F. Veinott, Jr., "Optimality of Myopic Inventory Policies for Several Substitute Products," Management Science *15*, 284-304 (1969).

[7] Nahmias, S., "Optimal Ordering Policies for Perishable Inventory—II," Operations Research *23*, 735-749 (1975).

[8] Nahmias, S., "On Ordering Perishable Inventory under Erlang Demand," Naval Research Logistics Quarterly *23*, 415-425 (1975).

[9] Nahmias, S., "Myopic Approximations for the Perishable Inventory Problem," Management Science *22*, 1002-1008 (1976).

[10] Nahmias, S. and W.P. Pierskalla, "A Two Product Perishable/Nonperishable Inventory Problem," SIAM Journal of Applied Mathematics *30*, 483-500 (1976).

[11] Van Zyl, G.J.J., "Inventory Control for Perishable Commodities," Unpublished Ph.D. dissertation, University of North Carolina, 1964.

[12] Veinott, Jr., A.F., "The Status of Mathematical Inventory Theory," Management Science *12*, 745-777 (1966).

NONLINEAR ONE-PARAMETRIC LINEAR PROGRAMMING AND T-NORM TRANSPORTATION PROBLEMS

Axel Wüstefeld and Uwe Zimmermann

University of Cologne
Federal Republic of Germany

ABSTRACT

Previous methods for solving the nonlinear one-parametric linear programming problem $\min \{c(t)^T x \mid Ax = b,\ x \geqslant 0\}$ for $t \in [\alpha,\beta]$ were based on the simplex method using a considerably extended tableau. The proposed method avoids such an extension. A finite sequence of feasible bases $(B_k \mid k = 1, 2, \ldots, r)$ optimal in $[t_k,t_{k+1}]$ for $k = 1, 2, \ldots, r$ with $\alpha = t_1 < t_2 < \ldots < t_{r+1} = \beta$ is determined using the zeroes of a set of nonlinear functions. Computational experience is discussed in the special case of t-norm transportation problems.

1. INTRODUCTION

Linear parametric programming has attracted many mathematicians as for example can be seen from the large reference list in Nožička et al. [3]. On the other hand there are only few results known for linear optimization problems with a nonlinear parameter. Carpentier [1], Sarkisjan [7] and Weinert [9] discuss linear optimization problems with a nonlinear parameter in the objective function. They consider the set of feasible solutions

$$P := \{x \in \mathbb{R}^n \mid Ax = b,\ x \geqslant 0\} \tag{1.1}$$

with real $m \times n$ matrix A of rank m and real positive m-vector b. For simplicity of the discussion it is always assumed that P is *nonempty, bounded and nondegenerate.* Let $I = [\alpha, \beta]$ with $\alpha,\beta \in \mathbb{R}$, $\alpha < \beta$ and let $c: I \to \mathbb{R}^n$ a continuous function. Unbounded intervals can be considered in the same way with minor changes in the assumptions. Then the linear optimization problem with nonlinear parameter in the objective function is

$$\min \{c(t)^T x \mid x \in P\} \qquad (t \in I). \tag{1.2}$$

In order to find for every $t \in I$ an optimal solution $x(t)$ it is sufficient to consider the finite set $\hat{P}$ of all bases corresponding to basic solutions of P. A finite sequence $(B_k \mid k = 1, 2, \ldots, r)$ of feasible bases B_k optimal in $t \in [t_k,t_{k+1}]$ for $k = 1, 2, \ldots, r$ with $\alpha = t_1 < t_2 < \ldots < t_{r+1} = \beta$ is called a *finite optimal solution of (1.2).*

In the above mentioned papers the problem is solved by means of the simplex method using an extended tableau. A certain transformation of the objective function in (1.2) leads to a new problem of the form

$$\min \{[d + D\ h(t)]^T x \mid x \in P\} \tag{1.3}$$

with a constant real vector d and a constant real $n \times k$ matrix D. Then (1.3) is discussed and

solved using results of linear parametric programming. The number of columns k is equal to the dimension of the smallest affine hyperspace containing $\{c(t) \mid t \in I\}$ (cf. [9]). The general linear optimization problem with one nonlinear parameter is

$$\min \{c(t)^T x \mid x \in P(t)\} \tag{1.4}$$

with the set of feasible solutions

$$P(t) := \{x \in \mathbb{R}^n \mid A(t)x = b(t),\ x \geqslant 0\} \tag{1.5}$$

defined by a real $m \times n$ continuous matrix function $A : I \to \mathbb{R}^{m \cdot n}$ of constant rank m and by a real positive continuous vector function $b: I \to \mathbb{R}^m$. If $P(t)$ is nonempty, bounded and nondegenerate for all $t \in I$ and if the set $\hat{P}(t)$ of all bases with respect to $t \in I$ is identical for all $t \in I$ then (1.4) can be solved in the same manner as (1.2). If all component functions are polynomials then a solution method can be found in Ritter [5]. In [5] the solution method is based on the optimality conditions of linear programming. This approach seems to be more direct and avoids the determination of the transformed problem (1.3).

In particular, we are interested in integer and combinatorial optimization problems having a linear characterization P such that the basic solutions of P are feasible solutions for the optimization problem. Then dependence on a continuously varying parameter t is considered only for the objective function. Thus, parametric programming in this case leads in a natural way to problems of the form (1.2).

Therefore, we develop a method for (1.2) based on the optimality conditions of linear programming in section 2 and discuss shortly the possible extension to (1.4) under the above assumptions. In section 3 the application of this method to transportation problems is discussed and t-norm objective functions that occur in algebraic transportation problems are treated (cf. [11]). The coefficient functions are of the form $f(t) = a^t$ with $a \in \mathbb{R}_+$. Finally some computational experience for t-norm transportation problems is given in section 4.

2. LINEAR PROGRAMMING WITH ONE NONLINEAR PARAMETER

As mentioned in the introduction we consider only the finite set of bases $\hat{P}$. A basis $B \in \hat{P}$ is the index vector of the *basic* variables. Then N denotes the index vector of the *nonbasic* variables. Partitions of vectors and matrices will be indexed by these vectors in the usual manner. The *set of all optimal bases* with respect to $t \in I$ is denoted by $V(t)$. We define

$$V(t',t'') := \bigcap_{t' \leqslant t \leqslant t''} V(t) \text{ for } t',\ t'' \in I \text{ with } t' \leqslant t''.$$

Then for $B \in \hat{P}$ the *optimality set* is $I(B) := \{t \in I \mid B \in V(t)\}$. This set can be characterized by means of the *reduced cost coefficient functions*

$$\bar{c}_j(t) := c_j(t) - c_B(t)^T A_B^{-1} A_j$$

for $j = 1, 2, \ldots, n$ with respect to a basis B. These functions are continuous in I. Obviously $\bar{c}_B(t) \equiv 0$. From linear programming we know

$$I(B) = \{t \in I \mid \bar{c}_N(t) \geqslant 0\}. \tag{2.1}$$

Thus the optimality set $I(B)$ is a closed subset of I. In the case of linear cost coefficient functions $c: I \to \mathbb{R}^n$ the parameter set $I(B)$ is a closed interval (cf. [3]). In general $I(B)$ may be disconnected (cf. [1], [9]). Then $I(B)$ is the union of mutually exclusive closed intervals called *optimality intervals of B*. For $t \in I(B)$ we denote the optimality interval containing t by $I_t^B =: [\alpha_t^B, \beta_t^B]$. We consider two sets of optimality intervals defined by

$$I^B := \{I_t^B \mid t \in I(B)\}$$

or $B \in \hat{P}$ and defined by

$$I_t := \{I_t^B \mid B \in V(t)\}$$

or $t \in I$. Obviously I_t is nonempty and finite for all $t \in I$. The existence of a finite solution of (1.2) is clearly equivalent to the existence of a *finite covering* $(I_{t_k}^{B_k} \mid k = 1, 2, \ldots, r)$ of I. Due to the compactness of I the following local characterization holds.

2.2) Local Existence Condition

For all $t \in (\alpha, \beta]$ resp. $t \in [\alpha, \beta)$ there exists $\tilde{B} \in \hat{P}$ resp. $B \in \hat{P}$ with

2.2.1) $\alpha_t^{\tilde{B}} < t \leqslant \beta_t^{\tilde{B}}$ resp.

2.2.2) $\alpha_t^B \leqslant t < \beta_t^B$.

(2.2) is necessary and sufficient for the existence of a finite optimal solution. Necessity is obvious and sufficiency follows by a compactness argument. We prefer to prove sufficiency in a constructive way. Let

2.3) $t_1 := \alpha$

$$t_k := \max \{\beta_{t_{k-1}}^B \mid B \in V(t_{k-1})\} \text{ for } k = 2, 3, \ldots$$

If (2.2) holds, then (2.3) yields a finite sequence $(t_k \mid k = 1, 2, \ldots, r+1)$ with $t_{r+1} = \beta$. Thus, this recursion provides a solution method for (1.2). To avoid the enumeration of $V(t)$, further conditions are discussed which are stronger than (2.2).

2.4) Finite Number of Optimality Intervals

Let I^B be finite for all $B \in \hat{P}$.

(2.4) implies (2.2). Furthermore, if (2.4) holds, then an arbitrary choice of the 'next' interval in (2.3) with

$$t_{k-1} < \beta_{t_{k-1}}^B$$

yields a finite sequence. The problem is to find such a basis $B \in V(t_{k-1})$. For a discussion of this problem we assume in the following $B \in V(\tau)$ for $\tau \in [\alpha, \beta)$. Then we have to check $< \beta_\tau^B$. As the optimality intervals will not be given explicitly, we must determine them with the aid of the reduced cost coefficients. (2.1) yields the following characterization:

2.5) $$\beta_\tau^B = \inf (\{t' \in (\tau, \beta] \mid \mathrm{E}i: \bar{c}_{N(i)}(t') < 0\} \cup \{B\}).$$

Numerical determination of the considered infimum seems to be very critical. It can easily be seen that β_τ^B is a zero for at least one of the nonbasic reduced cost coefficient functions. Therefore numerical methods for the determination of zeros of continuous functions can be applied. Unfortunately β_τ^B is in general not an isolated zero, which complicates the situation. Let

$$G_j^B := \{t \in I \mid \bar{c}_j(t) = 0\}$$

for $j = 1, 2, \ldots, n$ and $B \in \hat{P}$. If the following conditions holds, only isolated zeros are of interest.

2.6) Finite Number of Zeros

Let $G_j^B = I$ or let G_j^B be a finite set for all $j = 1, 2, \ldots, n$ and all $B \in \hat{P}$.

It is easily shown that (2.6) implies (2.4). Furthermore we can replace (2.5) by the determination of an isolated zero of the nonvanishing reduced cost coefficient functions $\bar{c}_j$, $j \in J(B) := \{i \mid G_i^B \neq I\}$:

$$\tilde{\tau}^{(B)} := \min \Big(\bigcup \{G_j^B \cap (\tau, \beta] \mid j \in J(B)\} \cup \{\beta\} \Big). \tag{2.7}$$

Now let $t' \in (\tau, \tilde{\tau})$. If $c_N(t') > 0$, then $[\tau, \tilde{\tau}]$ is a subset of an optimality interval of B. Otherwise, $\tau = \beta_\tau^B$ and we have to choose another solution $\tilde{B} \in V(\tau)$.

Even the calculation of the least zero (2.7) for a given set of at most $n - m$ nonvanishing nonbasic reduced cost coefficient functions remains very critical. No method is known to the authors that guarantees to find this very zero.

A further problem is to find a systematic way of changing the current basis in the case $\tau = \beta_\tau^B$. This is solved with the aid of the simplex method.

Therefore, we summarize in the following the changes due to a pivot step of the simplex method. The nonbasic variable $x_{N(s)}$ is exchanged for the basic variable $x_{B(r)}$. The new solution $\tilde{x}$ corresponds to the new basis $\tilde{B}$ defined by $\tilde{B}(i) = B(i)$ for all $i \neq r$ and $\tilde{B}(r) = N(s)$. $\tilde{N}$ is defined analogously. The new reduced cost coefficient functions are denoted by $\tilde{c}$. Furthermore, $a > o$ denotes the pivot element and e_r denotes the r-th unit vector of $\mathbf{R}^m$. Then from linear programming we know the following in parametric form for all $t \in I$:

$$c_{\tilde{B}}(t)^T A_{\tilde{B}}^{-1} = [c_B(t) + (\bar{c}_{N(s)}(t)/a)\, e_r]^T A_B^{-1} \tag{2.8.1}$$

$$\tilde{c}_{\tilde{N}}(t)^T = \bar{c}_{\tilde{N}}(t)^T - (\bar{c}_{N(s)}(t)/a)\, e_r^T A_B^{-1} A_{\tilde{N}} \tag{2.8.2}$$

$$\tilde{x}_{\tilde{B}(i)} = \begin{cases} x_{B(i)} - (x_{B(r)}/a)\,(A_B^{-1} A_{N(s)})_{B(i)} & i \neq r \\ x_{B(r)}/a & i = r. \end{cases} \tag{2.8.3}$$

Let $z(t)$ resp. $\tilde{z}(t)$ denote the values of the objective function in (1.2) with respect to B and $\tilde{B}$. Then

$$\tilde{z}(t) = z(t) + (x_{B(r)}/a) \cdot \bar{c}_{N(s)}(t) \tag{2.8.4}$$

for all $t \in I$. Due to the assumed nondegeneracy we know $x_{B(r)} > 0$.

Now we perform a sequence of pivot steps in the following way. If $B \notin V(\tau, \tilde{\tau})$ (cf. 2.7) then there exists a nonbasic variable $x_{N(s)}$ with

$$\bar{c}_{N(s)}(\tau) = 0 \tag{2.9.1}$$

$$\bar{c}_{N(s)}(t') < 0 \text{ for all } t' \in (\tau, \tilde{\tau}). \tag{2.9.2}$$

Then $x_{N(s)}$ is introduced into the basis. Due to (2.8.2) and (2.9.1) we find $\tilde{B} \in V(\tau)$. Subsequent pivot steps according to (2.9) have the following interpretation.

Let $\tau' := \min \{\tilde{\tau}(B') \mid B' \in V(\tau), \beta_\tau^{B'} = \tau\}$ with respect to (2.7). Then $\bar{c}_{N(s)}(t') < 0$ for all $t' \in (\tau, \tau')$ in each such pivot step. Therefore $z(t') > \tilde{z}(t')$ for all $t' \in (\tau, \tau')$. Thus a finite sequence of such pivot steps yields a basis $\tilde{B} \in V(\tau, \tau')$.

An analogous interpretation can be given assuming (2.4) instead of (2.6). Then $\tilde{\tau}$ in (2.9) is replaced by a parameter value $\tilde{\tilde{\tau}}_s$ defined by

$$\tilde{\tilde{\tau}}_s := \inf (\{t' \in (\tau, \beta] \mid \bar{c}_{N(s)}(t') \geqslant 0\} \cup \{\beta\}). \tag{2.10}$$

In the case $\tau = \beta_\tau^B$ there exists an index s with $\tilde{\tilde{\tau}}_s > \tau$. Then (2.9) holds and $x_{N(s)}$ is introduced into the basis.

We summarize the above proposed steps for a performance of (2.3) in the following algorithm. Finiteness and validity follow from the above remarks. The assumption (2.6) is made but if only (2.4) holds the necessary modifications can easily be found from (2.5) and (2.10).

(2.11) Solution Method For (1.2)

① (INITIAL SIMPLEX STEP)
Determine an initial basis $B \in V(\alpha)$; $t_1 := \tau := \alpha$; $k := 1$.

② (LEAST ZERO CALCULATION)
Determine $\tilde{\tau}$ according to (2.7);
if $\bar{c}_j(t') > 0$ for all $j \in J(B)$ and a parameter $t' \in (\tau, \tilde{\tau})$
then go to ④ .

③ (PIVOTING IN $V(\tau)$)
Choose a nonbasic variable $x_{N(s)}$ with $\bar{c}_{N(s)}(\tau) = 0$
and $\bar{c}_{N(s)}(t') < 0$ for a parameter $t' \in (\tau, \tilde{\tau})$;
perform a pivot step introducing $x_{N(s)}$ into the basis,
redefine x, B, N; go to ② .

④ (ITERATION)
$B^k := B$; $k := k + 1$;
if $\tilde{\tau} \geqslant \beta$ then $t_k := \beta$; $r := k$; stop.
Otherwise $t_k := \tau := \tilde{\tau}$; go to ② .

We have already emphasized the numerical difficulties in the determination of $\tilde{\tau}$ (cf. 2.7). Furthermore the necessary evaluations of the nonlinear functions $\bar{c}_N(t)$ will be highly time consuming in general (cf. section 4). Therefore simple numerical methods using only a few evaluations should be preferred. During the search in $V(\tau)$ the exact determination of $\tilde{\tau}$ can be replaced by any procedure giving an answer to the question whether the current basis B is optimal in a small interval $(\tau, \tau + \epsilon)$, $\epsilon > 0$ or not. Only if the answer is positive is an explicit determination of $\tilde{\tau}$ necessary.

In the following we discuss some theoretical properties of the solutions of (1.2) and its dual. Only the existence of a finite optimal solution of (1.2) is assumed.

We define the *optimal value function* $\hat{z} : I \to \mathbb{R}$ by

$$\hat{z}(t) := c_B(t)\, A_B^{-1} b \tag{2.12}$$

for $t \in I$ and $B \in V(t)$. This is a well defined function. Due to the existence of a finite solution, the continuity of $\hat{z}$ can be shown in the same way as in the case of linear functions $c(t)$.

The dual of (1.2) is

$$\max \{y^T b \mid y \in S(t)\} \tag{2.13}$$

with

$$S(t) := \{y \in \mathbb{R}^m \mid y^T A \leqslant c(t)^T\} \tag{2.14}$$

for $t \in I$. As (1.2) has an optimal solution for all $t \in I$ so has (2.13). A finite optimal solution

of (1.2) yields a *continuous optimal solution* $\hat{y} : I \to \mathbb{R}^m$ of (2.13) defined by

$$\hat{y}(t)^T := c_{B_k}(t)^T A_{B_k}^{-1}$$

for $t \in [t_k, t_{k+1}]$, $k = 1, 2, \ldots, r$. Continuity in t_k, $= 2, 3, \ldots, r$ follows from (2.8.1). The *optimal reduced cost coefficients* $\hat{c} : I \to \mathbb{R}^n_+$ defined by

$$\hat{c}_j(t) := c_j(t) - \hat{y}(t)^T A_j \quad \text{for } j = 1, 2, \ldots, n$$

are continuous functions in I.

If we consider the more general problem (1.4) the situation is only a little bit more complicated. This is due to the assumptions made for (1.4). For an application of the solution method only a few changes are necessary. The optimality set is now

(2.1)′ $$I(B) = \{t \in I \mid \bar{c}_N(t) \geqslant 0,\ x_B(t) \geqslant 0\}$$

with the reduced cost coefficients

$$\bar{c}_j(t) := c_j(t) - c_B(t)^T A_B^{-1}(t)\, A_j(t)$$

for $j = 1, 2, \ldots, n$. The finite optimal solution is not piecewise constant in general as before but is defined by

$$x_B(t) := A_B^{-1}(t) b(t)$$

in the respective intervals of I. A discussion of the pivot step changes in (2.8) shows that the dual optimal solution $\hat{y}$ and the optimal reduced cost functions $\hat{c}$ may be discontinuous at those optimality interval endpoints corresponding to the zeros of certain basic variables $x_{B(r)}(t)$. Then the primal optimal solution is continuous at such points. The necessary modifications of (2.11) in the definition of $\tilde{\tau}$ (resp. β_τ^B and $\tilde{\tilde{\tau}}_s$) follow easily from (2.1)′. The modified method solves problem (1.4) in the same manner as before (1.2).

3. TRANSPORTATION PROBLEMS AND T-NORM OBJECTIVES

The *nonlinear one-parametric transportation problem*

(3.1) $$\min_{x \in T} \sum_i \sum_j c_{ij}(t) \cdot x_{ij}$$

with a continuous function $c : I \to \mathbb{R}^{mn}$ subject to

(3.2) $$T := \{x \mid \sum_j x_{ij} = a_i,\ \sum_i x_{ij} = b_j,\ x_{ij} \geqslant 0\}$$

for given positive real numbers a_i, $i = 1, 2, \ldots, m$ and b_j, $j = m+1, m+2, \ldots, m+n$ is a special case of (1.2). W.l.o.g. we assume $\sum a_i = \sum b_j$. Index sets for i resp. j are given explicitly only if proper subsets of the general range $\{1, 2, \ldots, m\}$ resp. $\{m+1, m+2, \ldots, m+n\}$ are considered. The set of basic resp. nonbasic variable indices is denoted by B resp. N.

The primal transportation algorithm (cf. Murty [2]) is chosen as proper modification of the simplex method. It makes full use of the special structure of the transportation problem and provides a direct method for the construction of the nonbasic reduced cost coefficient functions $\bar{c}_{ij}(t)$ for $(i,j) \in N$. The dual variables are usually denoted by $u_i(t)$, $v_j(t)$. With respect to a given basis B they are recursively defined by

$$u_i(t) + v_j(t) = c_{ij}(t)$$

for all $(i,j) \in B$ starting with $u_1(t) \equiv 0$. The tree structure of B makes it easy to find the general form

$$f(t) = \sum_{B} \delta_{ij} \cdot c_{ij}(t)$$

with $\delta_{ij} \in \{0, \pm 1\}$ for all $(i,j) \in B$. Then the nonbasic reduced cost coefficient functions are given by

$$\bar{c}_{ij}(t) = c_{ij}(t) - u_i(t) - v_j(t)$$

for all $(i,j) \in N$. Therefore these functions are of the general form

$$\bar{c}_{\mu\nu}(t) = c_{\mu\nu}(t) - \sum_{B} \gamma_{ij} \cdot c_{ij}(t) \tag{3.3}$$

with $(\mu,\nu) \in N$ and $\gamma_{ij} \in \{0, \pm 1\}$ for all $(i,j) \in B$. Coefficients with value ± 2 do not occur as 'row nodes' μ have even distance and 'column nodes' ν have odd distance from the 'root node' 1 with respect to the tree B. For the same reason $\gamma_{ij} = 0$ for all (i,j) on the common 'path' from 'root node' 1 to the last common 'predecessor' of μ and ν. This information is helpful for the calculation of the reduced cost coefficients from the stored functions $c_{ij}(t)$. Only the storage of the iterated vectors γ of length $|B|$ is necessary. The storage required for the $c_{ij}(t)$ is decreased if the objective function has the form

$$\sum_{i} \sum_{j} \left[d_{ij}^{o} + \sum_{l=1}^{\bar{k}} g_l\,(d_{ij}^{l}, t) \right] x_{ij} \tag{3.4}$$

with certain functions g_l and constants d_{ij}^{l}, $l = 0, 1, \ldots, \bar{k}$. Weinert [9] and Carpentier [1] consider objective functions of the form

$$\sum_{i} \sum_{j} \left[d_{ij}^{o} + \sum_{l=1}^{\tilde{k}} h_l(t) \cdot d_{ij}^{l} \right] x_{ij} \tag{3.5}$$

with certain functions h_l and constants d_{ij}^{l}, $l = 0, 1, \ldots, \tilde{k}$. In (3.5) the coefficient functions depend only linearly on d_{ij}^{l}. This was necessary in their approach to define the rows of the extended simplex tableau (cf. (1.3)). The number of stored functions is only $\bar{k}$ resp. $\tilde{k}$ instead of mn. Furthermore $\bar{k} \leqslant \tilde{k}$.

A special case of (3.4) with $\bar{k} = 1$ is

$$z(x,t) : \sum_{i} \sum_{j} (c_{ij})^{t} \cdot x_{ij} \tag{3.6}$$

with positive real c_{ij}. For fixed $t \in [1, \infty)$ the minimization of (3.6) subject to (3.2) is equivalent to the *t-norm transportation problem*

$$z(t) := \min_{x \in T} \left[\sum_{i} \sum_{j} (c_{ij})^{t} x_{ij} \right]^{1/t}. \tag{3.7}$$

Such problems are examples for algebraic transportation problems (cf. [11]). For $t = 1$ (3.7) is the classical transportation problem. If we denote the objective function of (3.7) by $\tilde{z}(x,t)$ then it is easy to see that

$$\lim_{t \to \infty} \tilde{z}(x,t) = d(x) := \max \{c_{ij} \mid x_{ij} > 0\}$$

for all $x \in T$. This shows the close relationship of (3.7) with large values of the parameter and the *bottleneck transportation problem*

$$d := \min_{x \in T} d(x). \tag{3.8}$$

Further on we consider the *time-cost transportation problem*

$$(3.9)\qquad \begin{bmatrix} d \\ \delta \end{bmatrix} := \operatorname*{lex\,min}_{x \in T} \begin{bmatrix} d(x) \\ \sum_{c_{ij} = d(x)} x_{ij} \end{bmatrix}$$

that is a minimization with respect to the lexicographical ordering of vectors. Among the optimal solutions of (3.8) we determine those with minimum sum of the basic variables corresponding to the bottleneck. In order to exclude trivial cases we assume the existence of $c_{ij} < d$ and $d < c_{kl}$.

Then let

$$d_+ := \min \{c_{ij} \mid c_{ij} > d\}$$
$$d_- := \max \{c_{ij} \mid c_{ij} < d\}.$$

We assume the existence of a second-best solution of (3.9) with value $\delta' > \delta$ in the second component. Let ϵ denote the minimim value of a nonvanishing basic variable and let $V(t)$ resp. $\bar{V}$ resp. $\bar{\bar{V}}$ denote the set of optimal basic solutions for (3.7) resp. (3.8) resp. (3.9). Using an idea of Steinberg [8] we find the following relationships.

(3.10) PROPOSITION

$$\text{Let } \bar{\beta} := [\ln(\Sigma\, a_i) - \ln \epsilon] \,/\, [\ln d_+ - \ln d],$$
$$\bar{\bar{\beta}} := [\ln(\Sigma\, a_i - \delta) - \ln(\delta' - \delta)] / [\ln d - \ln d_-].$$

Then

$$(3.10.1)\qquad t > \bar{\beta} \Rightarrow V(t) \subseteq \bar{V},$$
$$t > \max(\bar{\beta}, \bar{\bar{\beta}}) \Rightarrow V(t) \subseteq \bar{\bar{V}}.$$

PROOF: Let $\bar{x} \in \bar{V}$ and $x \notin \bar{V}$. Then $t > \bar{\beta}$ implies
$z(\bar{x},t) \leqslant (\Sigma\, a_i) \cdot d' < \epsilon \cdot d'_+ \leqslant z(x,t)$.
Let $\bar{\bar{x}} \in \bar{\bar{V}}$ and $x \in \bar{V}, \bar{\bar{V}}$. Then $t > \max(\bar{\beta}, \bar{\bar{\beta}})$ implies
$z(\bar{\bar{x}},t) \leqslant \delta \cdot d' + (\Sigma\, a_i - \delta)\, d'_- < \delta' \cdot d' \leqslant z(x,t)$.

□

There are examples for which the bounds in (3.10) are sharp. Furthermore examples are known for which $\bar{\beta} < \bar{\bar{\beta}}$ and $\bar{\bar{V}} \cap V(t) = \phi$ for all t with $\bar{\beta} < t < \bar{\bar{\beta}}$.

For integer problems the bounds $\bar{\beta}$ and $\bar{\bar{\beta}}$ can be simplified using ϵ, $\delta' - \delta \geqslant 1$ and $d_+ \geqslant d + 1$, $d_- \leqslant d - 1$. (3.10) shows that with increasing parameters an optimal solution of (3.7) is necessarily an optimal solution of (3.8) and finally an optimal solution of (3.9). Due to these relations a parametric study of (3.7) or equivalently of (3.6) can be interpreted as discussion of various solutions 'between' the classical, the bottleneck, and the time-cost transportation problem. A detailed analysis can be given if (3.6) has a finite solution at least on the interval $[1, \max(\bar{\beta}, \bar{\bar{\beta}})]$.

With slightly modified assumptions (3.10) and the discussion of t-norm objectives in comparison with bottleneck-resp. time-cost objectives hold for the general problem (1.2). Furthermore a generalization for closed, bounded subsets P of $\mathbb{R}^n_+$ holds under similar assumptions.

Existence of finite solutions in the interval $[1, \max(\bar{\beta}, \bar{\bar{\beta}})]$ can be derived from the general form of the nonbasic reduced cost coefficients (3.3) for the unbounded interval $[1, \infty)$. In the special case of (3.6) these functions have the form $f(t) \equiv 0$ or

$$f(t) = \sum_{\delta=1}^{k} (s_\delta)^t - \sum_{\omega=1}^{l} (p_\omega)^t$$

with mutually distinct s_δ, $p_\omega > 0$ for $\delta = 1, 2, \ldots, k$ and $\omega = 1, 2, \ldots, l$. If $f(t) \not\equiv 0$ then $f(t)$ has at most $k + l - 1$ zeros in $\mathbb{R}$ (cf. Polya and Szegö [4]). Therefore (2.6) holds with respect to $[1, \infty)$. A bound for these zeros is (cf. Wüstefeld [10])

$$\tau \leqslant \ln(\max(k,l)) \,/\, [\ln \bar{\lambda} - \ln \underline{\lambda}] \tag{3.11}$$

with $\underline{\lambda} := \min\{\max s_\delta, \max p_\omega\} < \bar{\lambda} := \max\{\max s_\delta, \max p_\omega\}$. Such bounds are very useful in the calculation of the least zero $\tilde{\tau}$ in step ② of (2.11).

The optimal value function z of (3.7) has the following properties (cf. Wüstefeld [10]).

(3.12) PROPOSITION

Let $a_i \in \mathbb{R}$, $i = 1, 2, \ldots, m$ and $b_j \in \mathbb{N}$, $j = m+1, m+2, \ldots, m+n$. Then $z: [1, \infty) \to \mathbb{R}_+$ is a continuous, monotonically decreasing and piecewise differentiable function.

PROOF: The optimal solution function fulfills

$$z(t) = [c_{B_k}(t)^T x_{B_k}]^{1/t} \tag{3.13}$$

for $t \in [t_k, t_{k+1}]$, $k = 1, 2, \ldots, r$ with respect to a finite optimal solution of (3.6). Thus z is continuous and piecewise differentiable. Explicit differentiation in $t \in (t_k, t_{k+1})$ for $k \in \{1, 2, \ldots, r\}$ shows that the derivative is nonpositive iff

$$\sum_{B_k} (c_{ij})^t \cdot x_{ij} \cdot [\ln c_{ij} - \ln z(t)] \leqslant 0$$

with $(x_{ij}) := x_{B_k}$. Due to the integrality we know $x_{ij} \geqslant 1$ if $x_{ij} \neq 0$. Then $\ln c_{ij} \leqslant \ln z(t)$. Together with (3.12) this yields the monotonicity of z. □

Obviously $z(t)$ is monotonically decreasing with limit d. For sufficiently large t (cf 3.11) we find a constant optimal solution for (3.7) which is optimal for (3.8) and (3.9) as well.

4. COMPUTATIONAL EXPERIENCE

The method discussed in the previous sections was coded for t-norm transportation problems with $t \in [1, \beta]$, $\beta := \ln(\Sigma a_i) / [\ln(d+1) - \ln d]$ (cf. 3.10), in FORTRAN IV. Tests were run on the CDC CYBER 72/76 at the computer center of the University of Cologne.

The random integer coefficients of the nxm cost matrices were drawn from a uniform distribution in the interval (0,100) resp. from a (50,3) Gauss distribution. The integers a_i and b_j were chosen in (1,10) with $\Sigma a_i = \Sigma b_j$. Twenty-five examples were solved for each choice of the dimension parameters n, m.

No significant differences were found between quadratic ($n = m$) and rectangular problems ($n \neq m$). Therefore we discuss only the case $n = m$. In comparison with transportation problems without parameters, computing times were raised enormously. Evaluation of the functions $\bar{c}_{\mu\nu}(t)$ in the determination of zeros (step 2 of 2.11) consumed about 70% of the computing times.

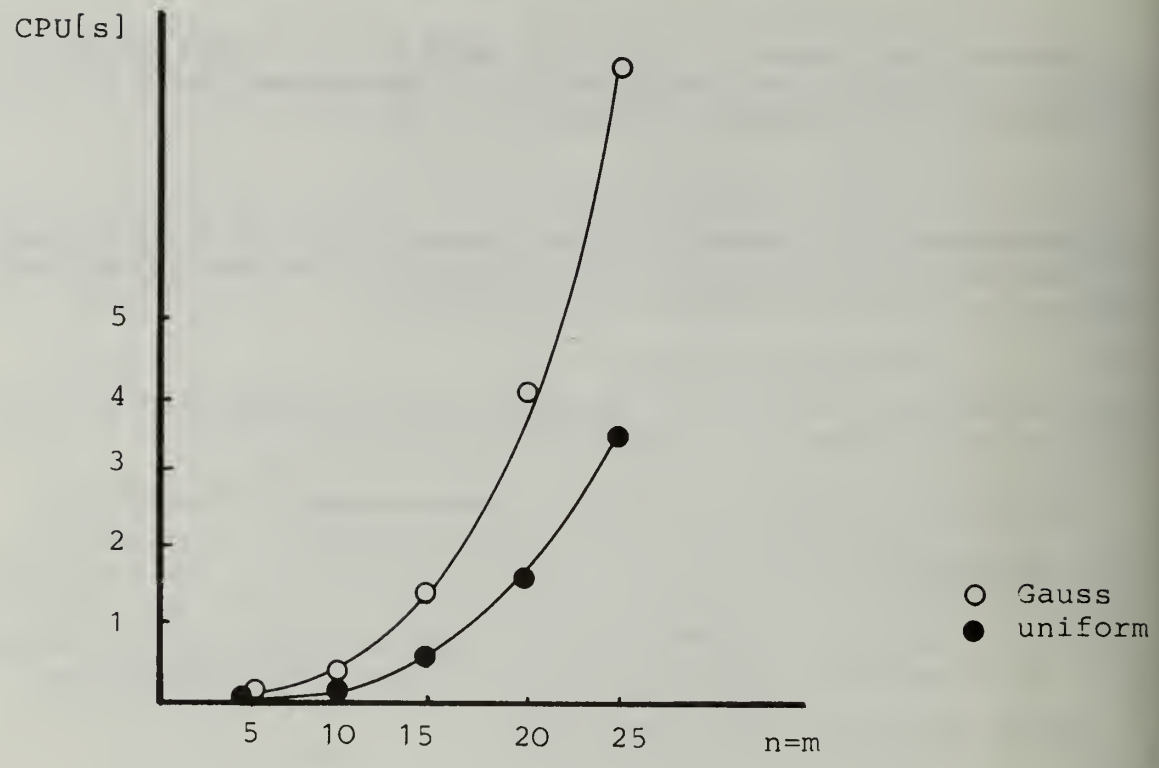

FIGURE 1.

Problems with Gaussian distributed random integer cost coefficients show to be more difficult. This is essentially due to less efficient bounds (cf. 3.11) in this case. Then more function evaluations are used during the performance of the bisection method which was implemented for the determination of zeros.

The mean running time is approximately of order $0(n^4)$. This coincides with the product of the number of the reduced cost coefficient functions $(n-1)^2$ and the approximate order $0(n^2)$ of the number of pivotings.

TABLE 1—*Mean Number of Pivoting Steps for T-Norm Transportation Problems in [1, β]*

n = m	5	10	15	20	25
uniform	2.	4.4	8.8	13.4	16.2
Gauss	1.8	5.2	8.5	13.6	19.1

Standard deviations of both considered means are rather high in all cases (30-50% of the mean values). Obviously the practical solvability of parametric problems is limited by running times. Storage problems are less important. In the proposed method, storage is essentially given by the size of the matrix of the cost coefficients. Compared with the simplex method using an extended tableau (cf. [9]) this is an obvious advantage.

BIBLIOGRAPHY

[1] Carpentier, J., "A Method for Solving Linear Programming Problems in Which the Cost Depends Nonlinearly on a Parameter," Électricité de France, Diréction de Etudes et Recherches (May 1959).

[2] Murty, K.G., *Linear and Combinatorial Programming* (Wiley and sons, New York 1976).
[3] Nožička, F., J. Guddat, H. Hollatz and B. Bank, "Theorie der Linearen Parametrischen Optimierung," Akademie-Verlag, Berlin (1974).
[4] Polya, G. and G. Szegö, "Problems and Theorems in Analysis II," Springer-Verlag, Berlin (1976).
[5] Ritter, K., "A Method for Solving Nonlinear Maximum-Problems Depending on Parameters," Naval Research Logistics Quarterly 14, 147-162 (1967).
[6] Ritter, K., Über Probleme Parameterabhängiger Planungsrechnung (DVL-Bericht Nr. 238)," Vereinigte Universitäts und Fachbuchhandlungen, O. Müller-Trewendt & Garnier-H. Sach GmbH, Köln (1963).
[7] Sarkisjan, S.D., "On a Parametric Linear Optimization Problem with Nonlinear Parameter Dependent Objective," Trudy vyčislit. Centra Akad. Nauk Armjan SSR. Erevan gosudarst. Univ. 2, 10-16 (Russian) (1964).
[8] Steinberg, L., "The Backboard Wiring Problem: A Placement Algorithm," Society for Industrial and Applied Mathematics Review 3, 37-50 (1961).
[9] Weinert, H., "Probleme der Linearen Optimierung mit Nicht linear-Einparametrischen Koeffizienten in der Zielfunktion," Mathematische Operationsforschung und Statistik 1, 21-43 (1970).
[10] Wüstefeld, A., "Einparametrische Optimierung für die P-Norm Zielfunktion beim Transportproblem," Diplomarbeit, Math. Inst. Univ. Köln, Germany (1978).
[11] Zimmermann, U., "Duality Principles and the Algebraic Transportation Problem," in Numerische Methoden bei graphen theoretischen und kombinatorischev Problemen, Band 2, ed. by L. Collatz, G. Meinardus and W. Wetterling, Birkhäuser Verlag Basel, Boston, Stuttgart, ISNM 46 (1979) 234-255.

OPTIMAL LOCATION OF A FACILITY RELATIVE TO AREA DEMANDS*

Z. Drezner

Faculty of Management
University of Michigan (Dearborn)

G. O. Weslowsky

McMaster University
Hamilton, Ontario, Canada

ABSTRACT

This paper deals with the Weber single-facility location problem where the demands are not only points but may be areas as well. It provides an iterative procedure for solving the problem with l_p distances when $p > 1$ (a method of obtaining the exact solution when $p = 1$ and distances are thus rectangular already exists). The special case where the weight densities in the areas are uniform and the areas are rectangles or circles results in a modified iterative process that is computationally much faster. This method can be extended to the simultaneous location of several facilities.

INTRODUCTION

In the ordinary Weber problem, we must locate a facility so that the sum of weighted distances from the facility to n demand points is minimized. There are cases where the demand may be better considered spread over an area or areas than concentrated at mathematical points on the plane. For example, one could be dealing with a very large number of demand points as in an urban area. Then too, a "distributed" demand could be used to represent random occurrences of demand from within areas.

When distances are rectangular (l_1) and when the areas are themselves rectangles with sides parallel or perpendicular to the axes of measurement, the problem can be solved exactly (Wesolowsky and Love [7]). The problem becomes more difficult with Euclidean (l_2) distances. Love [4] expressed the objective function in analytic terms and suggested a nonlinear optimization procedure. Bennett and Mirakhor [2] replaced the areas with their centers of gravity as an approximation.

Our approach is essentially based on the Weiszfeld [6] procedure; that is, on an iterative solution of the extremal equations. In the special case when areas are rectangles (as assumed by [2], [4], and [7]) or circles, and demand is uniformly distributed, we propose a "special" iterative procedure.

Our method is more general than the methods in [2] and [4] because l_p distances and not only Euclidean distances are permitted. Also, circles and general shapes are treated in addition to rectangles. Further, our method appears considerably faster than Love's method and gives more accurate results than Bennett and Mirakhor's approach.

*This research was supported by the Natural Sciences and Engineering Research Council Canada.

THE PROBLEM

Consider the distribution of weight density over a plane. The weight density could be interpreted as the cost per unit distance per unit area (if the area is infinitesimally small) of delivering to the demand in that area. We wish to find the location of a facility (x_0,y_0) that minimizes the total cost of delivery. We must therefore minimize

$$F(x_0,y_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\,[|x-x_0|^p + |y-y_0|^p]^{1/p}\,dxdy, \tag{1}$$

where $w(x,y)$ is the weight density, $w(x,y) \geq 0$, and $[|x-x_0|^p + |y-y_0|^p]^{1/p}$, $p \geq 1$, is the l_p distance of point (x,y) from (x_0,y_0). Note that for $p = 1$, distances are rectangular and that for $p = 2$ they are Euclidean.

We will consider only $p > 1$ because a simple procedure for solving the problem when $p = 1$ already exists [7].

It is safe to assume that in practice we can find a finite circle large enough so that $w(x,y)$ is zero outside it. It will be helpful to consider the following discrete approximation to Eq. (1); we use it only to make certain provable conclusions about Eq. (1) intuitive.

$$f(x_0,y_0) = \sum_{j=1}^{n} w_j[|x_j-x_0|^p + |y_j-y_0|^p]^{1/p}\,dxdy. \tag{2}$$

The function $f(x_0,y_0)$ is simply the objective function of the ordinary Weber problem with n demand points. It is clear that we can choose the weights w_j and the number n to make $f(x_0,y_0)$ approximate $F(x_0,y_0)$ as closely as desired. We could thus think of $F(x_0,y_0)$ as the limit of $f(x_0,y_0)$ as $n \to \infty$.

Since $f(x_0,y_0)$ is known to be convex (it is a nonnegative linear combination of convex distance norms), so is $F(x_0,y_0)$ (this could be shown formally). The minimum is therefore defined by the extremal conditions

$$\frac{\partial F}{\partial x_0} = \frac{\partial F}{\partial y_0} = 0.$$

Note that the derivatives of $f(x_0,y_0)$ do not exist at $(x_0,y_0) = (x_j,y_j)$. This problem does not occur with $F(x_0,y_0)$. Differentiating under the integral we have

$$\frac{\partial F}{\partial x_0} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\,\frac{(x-x_0)|x-x_0|^{p-2}}{[|x-x_0|^p + |y-y_0|^p]^{1-1/p}}\,dxdy = 0, \tag{3a}$$

$$\frac{\partial F}{\partial y_0} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(x,y)\,\frac{(y-y_0)|y-y_0|^{p-2}}{[|x-x_0|^p + |y-y_0|^p]^{1-1/p}}\,dxdy = 0. \tag{3b}$$

These equations can be used to construct an iterative procedure that is similar to the Weiszfeld procedure [6] and was suggested in [5]. In effect, we "solve" Eqs. (3a) and (3b) and use the result iteratively. For example, from Eq. (3a) we obtain

$$x_0^{(k+1)} = \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{w(x,y)\,x|x-x_0^{(k)}|^{p-2}}{[|x-x_0^{(k)}|^p + |y-y_0^{(k)}|^p]^{1-1/p}}\,dxdy}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{w(x,y)|x-x_0^{(k)}|^{p-2}}{[|x-x_0^{(k)}|^p + |y-y_0^{(k)}|^p]^{1-1/p}}\,dxdy}, \tag{4}$$

where k stands for the k th iteration. A similar expression can be obtained for $y_0^{(k+1)}$. This iterative procedure (for the discrete case $f(x_0,y_0)$ and thus for $F(x_0,y_0)$) is known to converge for $p = 2$. It has not been proven to converge for $p \neq 2$, but to our knowledge has always converged in practice [5]. When it does converge, however, the optimality conditions are satisfied and the optimum has been reached.

As an illustration, let $w(x,y) = \sqrt{x^2 + y^2}$ only inside the unit circle around the origin. If $p = 2$, Eq. (4) becomes

$$x_0^{(k+1)} = \frac{\int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{x\sqrt{x^2+y^2}}{\sqrt{(x - x_0^{(k)})^2 + (y - y_0^{(k)})^2}}\, dydx}{\int_{-1}^{1} \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{\sqrt{x^2+y^2}}{\sqrt{(x - x_0^{(k)})^2 + (y - y_0^{(k)})^2}}\, dydx}.$$

Iterations will produce convergence to the origin. For example, when $(x_0^{(k)},y_0^{(k)})$ is near the origin it can be shown that $x_0^{(k+1)} \approx x_0^{(k)}/2$ and $y_0^{(k+1)} \approx y_0^{(k)}/2$.

The numerical computation of double integrals in Eq. (4) could consume a great deal of time. We suggest a more efficient two-stage procedure in the next section.

A TWO-STAGE ITERATIVE PROCEDURE

The total area of demand is usually made up of distinct areas where $w(x,y) = 0$ between them. In any case, without a loss of generality, we can break up the total area into q subareas, $A_1, \ldots, A_q$. We ignore for the moment the possible existence of demands at mathematical points to simplify the presentation; such points can be easily included into consideration later.

Under these assumptions, Eqs. (3a) and (3b) become

$$\frac{\partial F}{\partial x_0} = -\sum_{i=1}^{q} \iint_{A_i} \frac{w(x,y)\,(x - x_0)\,|x - x_0|^{p-2}}{[|x - x_0|^p + |y - y_0|^p]^{1-1/p}}\, dxdy,$$

(5a)
$$\frac{\partial F}{\partial x_0} = -\sum_{i=1}^{q} F_{xi}(x_0,y_0) = 0$$

$$\frac{\partial F}{\partial y_0} = -\sum_{i=1}^{q} \iint_{A_i} \frac{w(x,y)\,(y - y_0)\,|y - y_0|^{p-2}}{[|x - x_0|^p + |y - y_0|^p]^{1-1/p}}\, dxdy$$

(5b)
$$\frac{\partial F}{\partial y_0} = -\sum_{i=1}^{q} F_{yi}(x_0,y_0) = 0.$$

The central idea behind the proposed iterative procedure is that *given* a location (x_0,y_0) of the facility, the cost of area i can be replaced with the cost of an equivalent demand point whose location and weight can be computed. In effect, we will then be using a Weiszfeld procedure on q point destinations.

Consider $(\hat{x},\hat{y})$, which can be any point on the plane and any area A_i, such that $F_{xi}(\hat{x},\hat{y}) \neq 0$ and $F_{yi}(\hat{x},\hat{y}) \neq 0$. We shall find (x_i,y_i) and w_i', the location and weight of a point which will give derivative components equal to $F_{xi}(\hat{x},\hat{y})$ and $F_{yi}(\hat{x},\hat{y})$. In effect, we are replacing the area A_i with an "equivalent" point; note, however, that this equivalence is dependent on the location $(\hat{x},\hat{y})$ and is not necessarily true elsewhere. It follows that at optimality, q points (x_i,y_i) with weights w_i' can replace the problem with areas. Although (x_i,y_i) and w_i' are functions of $(\hat{x},\hat{y})$, we do not specify this with notation to avoid unnecessary notational complexity.

Therefore, by Eqs. (5a) and (5b),

$$(6a) \qquad w_i' \frac{(x_i - \hat{x})|x_i - \hat{x}|^{p-2}}{[|x_i - \hat{x}|^p + |y_i - \hat{y}|^p]^{1/p}} = F_{xi}(\hat{x},\hat{y})$$

$$(6b) \qquad w_i' \frac{(y_i - \hat{y})|y_i - \hat{y}|^{p-2}}{[|x_i - \hat{x}|^p + |y_i - \hat{y}|^p]^{1/p}} = F_{yi}(\hat{x},\hat{y}).$$

Note that the left-hand sides in Eqs. (6a) and (6b) are simply the derivatives of $w_i'[|x_i - \hat{x}|^p + |y_i - \hat{y}|^p]^{1/p}$ with respect to x and y, respectively, and the right-hand sides have the corresponding derivatives for the term A_i.

Dividing Eq. (6a) by Eq. (6b), we get

$$(7) \qquad \left(\frac{x_i - \hat{x}}{y_i - \hat{y}}\right) \frac{|x_i - \hat{x}|^{p-2}}{|y_i - \hat{y}|^{p-2}} = \frac{F_{xi}(\hat{x},\hat{y})}{F_{yi}(\hat{x},\hat{y})} = R_i(\hat{x},\hat{y}).$$

Let

$$s_i(\hat{x},\hat{y}) = 1, \quad \text{if } R_i(\hat{x},\hat{y}) > 0,$$
$$= -1, \quad \text{if } R_i(\hat{x},\hat{y}) < 0.$$

Therefore, Eq. (7) implies

$$(8) \qquad y_i = (x_i - \hat{x})/(s_i(\hat{x},\hat{y})|R_i(\hat{x},\hat{y})|^{1/(p-1)}) + \hat{y}$$

which is the equation of a straight line through $(\hat{x},\hat{y})$; in other words, (x_i,y_i) is not uniquely determined. Some comments on technical points should be made. When $(\hat{x},\hat{y})$ is such that $F_{xi} = 0$ and $F_{yi} \neq 0$, then $\hat{x}_i = x_i$ and y_i is undetermined. If $F_{yi} = 0$, but $F_{xi} \neq 0$, then $y_i = \hat{y}$ and x_i is undetermined. If both $F_{xi} = 0$ and $F_{yi} = 0$, then $(x_i,y_i) = (\hat{x},\hat{y})$. Further (barring $F_{xi} = 0$ and $F_{yi} = 0$), when $x_i = \hat{x}$ and $y_i = \hat{y}$, then Eq. (8) still gives the equation of the line in spite of the fact that the left-hand side of Eq. (7) is undefined. The above comments could be proved analytically, but it is easier to justify them (and, in fact, the rather surprising result of Eq. (8) as well) by invoking a mechanical analog that has been used for solving the Weber problem with point demands ([3], p. 193). First, however, we complete our assignment by finding w_i', the equivalent weight of the point.

We take absolute values on both sides of Eqs. (6a) and (6b), raise both sides to the power $p/(p-1)$, and sum the two equations to obtain

$$(w_i')^{p/p-1} \frac{|x_i - \hat{x}|^p + |y_i - \hat{y}|^p}{[|x_i - \hat{x}|^p + |y_i - \hat{y}|^p]} = |F_{xi}|^{p/(p-1)} + |F_{yi}|^{p/(p-1)}.$$

Therefore,

$$(9) \qquad w_i' = [|F_{xi}|^{p/(p-1)} + |F_{yi}|^{p/(p-1)}]^{(p-1)/p}.$$

We now turn to the Varignon frame, a mechanical frame analog, for an explanation of the results Eqs. (8) and (9). The frame is constructed by drilling holes in a board corresponding to the coordinates of demand points. Weights corresponding to weights w_i are hung from strings which are passed through the board from below and tied together in a knot above. It is easily demonstrated that the knot is in equilibrium at the best location for the facility when $p = 2$. However, consider the knot at some point $(\hat{x},\hat{y})$ other than at equilibrium. Note first that the force exerted on the knot by any weight depends on the magnitude of the weight and the angle of the string, not on the distance of the weight from the knot. This "explains" the line relationship (8) and the indeterminancy of the exact position of w_i'. Further, we could imagine the

areas as densely filled with holes and having *very* many strings leading to the knot. The condition $F_{xi}(\hat{x},\hat{y}) = 0$, $F_{yi}(\hat{x},\hat{y}) = 0$ is equivalent to the knot being in such a position that the forces of the area strings balance; this explains why w_i' is zero here (as can be checked by Eq. (9)). Also, when, for example, only $F_{xi}(\hat{x},\hat{y}) = 0$, this means that there is no X-direction component exerted by the strings, and Eq. (8) is replaced by a vertical line through $\hat{x}$.

The iterative procedure we now suggest is as follows. We start with some initial point $(\hat{x}^0,\hat{y}^0)$. We then convert the areas A_i into points with weights w_i' (if A_i is a point, $w_i' = w_i$). Since the exact location of a point on the line (8) is indeterminate, we choose that point that is closest to the center point of the rectangle bounding A_i. We now use the Weiszfeld iteration for points [5]:

$$\hat{x}^{(k+1)} = \frac{\sum_{i=1}^{q} \dfrac{w_i' x_i |x_i - \hat{x}^{(k)}|^{p-2}}{[|x_i - \hat{x}^{(k)}|^p + |y_i - \hat{y}^{(k)}|^p]^{1-1/p}}}{\sum_{i=1}^{q} \dfrac{w_i' |x_i - \hat{x}^{(k)}|^{p-2}}{[|x_i - \hat{x}^{(k)}|^p + |y_i - \hat{y}^{(k)}|^p]^{1-1/p}}}. \tag{10a}$$

The equation for $\hat{y}^{(k+1)}$ is analogous and will be referred to as Eq. (10b). This procedure could be modified to a hyperbolic form as in [5]. The new point is used to recalculate equivalent weights, then $\hat{x}^{(k+1)}$ and $\hat{y}^{(k+1)}$ are recalculated, and so on.

We do not have a proof that the above procedure converges. However, we have found no cases where Eq. (10) did not converge very quickly. Also, note that when Eq. (10) does converge, the optimum has been found because when the left-hand side of Eq. (10) is equal to the right-hand side, the extremal equations are met. The advantage of Eq. (10) over Eq. (4) is that for certain special cases, we can find relatively simple forms for the weight w_i' and for Eq. (8).

The method can be extended to a problem as in [7] where several facilities (with intershipments) are to be located. The Weiszfeld procedure can be adapted to the multifacility problem along the lines described on p. 230 of [3]. Since only the demands are areas, Eq. (9) could be used in the same way. In the next section, we will deal with the problem under some simplifying assumptions which, incidentially, have also been made by previous papers on the subject.

SPECIAL ASSUMPTIONS

We first assume that the weight density in each area A_i is uniform and we denote it by w_i^0. We now have

$$F_{xi}(\hat{x},\hat{y}) = w_i^0 \iint_{A_i} \frac{(x - \hat{x})|x - \hat{x}|^{p-2}}{[|x - \hat{x}|^p + |y - \hat{y}|^p]^{1-1/p}}\, dxdy. \tag{11}$$

If we define A_i by letting y be in the region $[\phi_{1i}(x), \phi_{2i}(x)]$ for $a_i \leqslant x \leqslant b_i$ and by letting x be in the region $[\psi_{1i}, (y), \psi_{2i}(y)]$ for $c_i \leqslant y \leqslant d_i$,

$$F_{xi}(\hat{x},\hat{y}) = w_i^0 \int_{c_i}^{d_i} [|x - \hat{x}|^p + |y - \hat{y}|^p]^{1/p} \Big|_{\psi_{1i}(y)}^{\psi_{2i}(y)} dy$$

and therefore

$$F_{xi}(\hat{x},\hat{y}) = w_i^0 \int_{c_i}^{d_i} \{[|\psi_{2i}(y) - \hat{x}|^p + |y - \hat{y}|^p]^{1/p} - [|\psi_{1i}(y) - \hat{x}|^p + |y - \hat{y}|^p]^{1/p}\} dy. \tag{12}$$

Similarly,

$$F_{yi}(\hat{x},\hat{y}) = w_i^0 \int_{a_i}^{b_i} \{[|x - \hat{x}|^p + |\phi_{2i}(x) - \hat{y}|^p]^{1/p} - [|x - \hat{x}|^p + |\phi_{1i}(x) - \hat{y}|^p]^{1/p}\} dx. \tag{13}$$

The above one-dimensional definite integrals can be evaluated by numerical integration. Note that Eqs. (12) and (13) may not be usable in that form if A_i is not a convex set; however, integration could then be done in "pieces".

We now look at two special shapes of areas when distances are Euclidean ($p = 2$): circles and rectangles. First, we consider finding the equivalent direction of a circular area. It could be shown that this line of force always passes through the center of the circle. This is obvious by consulting the mechanical analog. The strings emanating from the much drilled circular area must pull on the knot at $(\hat{x},\hat{y})$ in such a way that the resultant force passes through the center (recall that equal weight density is assumed). To compute the weight w_i', we consider the point $(\hat{x},\hat{y}) = (0,0)$ and a circle of radius R centered at $(a,0)$. It follows that $F_{yi} = 0$. Note that from symmetry considerations the weight obtained is the same as the weight of any circle of radius R with its center "a" units from any $(\hat{x},\hat{y})$.

By Eqs. (9) and (12) and assuming $p = 2$,

$$\begin{aligned} w_i' &= |F_{xi}| \\ &= w_i^0 \int_{-R}^{R} \{[(\sqrt{R^2 - y^2} + a)^2 + y^2]^{1/2} - [(\sqrt{R^2 - y^2} - a)^2 + y^2]^{1/2}\} dy \\ &= w_i^0 R^2 \int_{-1}^{1} \{[(\sqrt{1 - y^2} + a/R)^2 + y^2]^{1/2} - [(\sqrt{1 - y^2} - a/R)^2 + y^2]^{1/2}\} dy \end{aligned}$$

$$w_i' = w_i^0 \pi R^2 f(a/R) \tag{14}$$

where

$$f(\alpha) = \frac{1}{\pi} \int_{-1}^{1} \{[(\sqrt{1 - y^2} + \alpha)^2 + y^2]^{1/2} - [(\sqrt{1 - y^2} - \alpha)^2 + y^2]^{1/2}\} dy.$$

The evaluation of $f(\alpha)$ can be done accurately by the evaluation of truncated series as described in the appendix.

We now turn to the rectangle shape that was also treated in [2], [4], and [7]. In Eqs. (12) and (13), we will have to use

$$\psi_{1i}(y) = a_i, \quad \psi_{2i}(y) = b_i$$
$$\phi_{1i}(x) = c_i, \quad \phi_{2i}(x) = d_i$$

giving

$$F_{xi}(\hat{x},\hat{y}) = w_i \int_{c_i}^{d_i} \{[(b_i - \hat{x})^2 + (y - \hat{y})^2]^{1/2} - [(a_i - \hat{x})^2 + (y - \hat{y}^2)^2]^{1/2}\} dy$$

and a similar expression for $F_{yi}(\hat{x},\hat{y})$.

These expressions can be evaluated to give

$$\begin{aligned} F_{xi} = w_i^0 \Bigg\{ & (\hat{x} - b_i)|\hat{x} - b_i| \left[f\left(\frac{\hat{y} - d_i}{\hat{x} - b_i}\right) - f\left(\frac{\hat{y} - c_i}{\hat{x} - b_i}\right) \right] \\ & - (\hat{x} - a_i)|\hat{x} - a_i| \left[f\left(\frac{\hat{y} - d_i}{\hat{x} - a_i}\right) - f\left(\frac{\hat{y} - c_i}{\hat{x} - d_i}\right) \right] \Bigg\} \end{aligned} \tag{15a}$$

$$\begin{aligned} F_{yi} = w_i^0 \Bigg\{ & (\hat{y} - d_i)|\hat{y} - d_i| \left[f\left(\frac{\hat{x} - b_i}{\hat{x} - d_i}\right) - f\left(\frac{\hat{x} - a_i}{\hat{y} - d_i}\right) \right] \\ & - (\hat{y} - c_i)|\hat{y} - c_i| \left[f\left(\frac{\hat{x} - b_i}{\hat{x} - a_i}\right) - f\left(\frac{\hat{y} - c_i}{\hat{x} - a_i}\right) \right] \Bigg\} \end{aligned} \tag{15b}$$

where

$$f(\alpha) = [\alpha \sqrt{1+\alpha^2} + ln(\alpha + \sqrt{1+\alpha^2})].$$

COMPUTATIONAL RESULTS AND CONCLUSIONS

Test problems with points (for comparison) and areas in the shape circles, rectangles, and ellipses were run. The points and centers of areas were randomly selected from within a 10 × 10 area. Table 1 gives the computational results. The procedure was stopped when the l_2 distance between two successive iterations was less than 10^{-4}. The considerably longer time taken for ellipses was due not only to numerical integration, but because p was 1.78 and computing quantities to a power other than 2 takes longer.

TABLE 1 — *Computational Results*

No. Demands	Type of Demand	No. Iterations	CDC 6400 s	Distance
500	points	11	0.86	l_2
500	rectangles	6	6.23	l_2
500	circles	6	1.02	l_2
50	ellipses	11	41.9	$l_{1.78}$

The unconstrained example given by Love in [4] was solved in 0.1 s as opposed to the 21 s on the Burroughs B-5500 reported by Love; the same solution was obtained but the convergence criteria were on different principles. As a very rough guess the CDC 6400 is 5 or 6 times faster.

It should be noted that the approximation suggested by Bennett and Mirakhor [2], namely, replacing areas by their centroids, could give widely different results in some examples. For example, consider the two rectangular areas in Table 2. The optimum point is (1.32337,0) with a cost of 41.15065. If we replace the areas with centroids, the solution is (5.5,0) with a cost of 58.58114.

TABLE 2 — *A Two-Area Example*

i	w_i	a_i	b_i	c_i	d_i
1	8	0	1	-0.5	0.5
2	0.9	1	10	-0.5	0.5

BIBLIOGRAPHY

[1] Abramowitz, M. and I.A. Stegum, eds., *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematics Series 55, pp. 589-607.

[2] Bennett, C.D. and A. Mirakhor, "Optimal Facility Location with Respect to Several Regions," Journal of Regional Science *14*, pp. 131-136 (1974).

[3] Francis, R.L. and J. White, *Facility Layout and Location*, Prentice-Hall, Inc. (1974).

[4] Love, R.F., "A Computational Procedure for Optimally Locating a Facility with Respect to Several Rectangular Regions," Journal of Regional Science *12*, pp. 233-242 (1972).

[5] Morris, J.G. and W.A. Verdini, "A Simple Iterative Scheme for Solving Minisum Facility Location Problems Involving l_p Distances," Wisconsin Working Paper 8-77-31, Graduate School of Business, University of Wisconsin-Madison.

[6] Weiszfeld, E., "Sur le point pour lequel la somme des distances de n points donnes est minimum," Tohoku Mathematics Journal *43*, pp. 355-386 (1936).

[7] Wesolowsky, G.O. and R.F. Love, "Location of Facilities with Rectangular Distances Among Point and Area Destinations," Naval Research Logistics Quarterly *18*, pp. 83-90 (1971).

Appendix
EVALUATION OF f(α)

As given in Eq. (14),

$$f(\alpha) = \frac{1}{\pi} \int_{-1}^{1} \left\{ \left[\left(\sqrt{1-y^2} + \alpha \right)^2 + y^2 \right]^{1/2} - \left[\left(\sqrt{1-y^2} - \alpha \right)^2 + y^2 \right]^{1/2} \right\} dy.$$

This is an elliptic integral (see [1]) with the following properties:

$$f(0) = 0,\ f(\infty) = 1,$$

$$f(1) = 8/3\pi,\ f'(\alpha) > 0 \text{ for } \alpha > 0,$$

$$f(\alpha) = \alpha f(1/\alpha).$$

We can compute $f(\alpha)$ by numerical integration.

In standard tables, we can find the series formulation

$$f(\alpha) = \frac{\sqrt{1+\alpha^2}}{2} \sum_{k=0}^{\infty} \left(\frac{2\alpha}{1+\alpha^2} \right)^{2k+1} \frac{(4k)!}{2^{6k}(2k)!(k!)^2(k+1)}.$$

Since convergence is slow for $\alpha \approx 1$, we can obtain an accelerated formula by defining

$$\beta = 2\alpha/(1+\alpha^2),$$

$$a_k = [\sqrt{2}/2\pi k - (4k)!/(2^{6k}(2k)!(k!)^2)]/(k+1),$$

$$f(\alpha) = \sqrt{1+\alpha^2}/2 \left\{ \beta + (\sqrt{2}/2\pi)\ [\beta + ((1-\beta^2)/\beta) ln(1-\beta^2)] - \sum_{k=1}^{\infty} a_k \beta^{2k+1} \right\}.$$

The function $f(\alpha)$ can be evaluated approximately as follows:

(a) for $\alpha \in [0,1]$,

(i) $0 \leqslant \alpha \leqslant 0.5$,

$$f(\alpha) = 2\alpha \frac{4-\alpha^2}{8-\alpha^2},$$

to a maximum error of 3×10^{-5};

(ii) for $0.5 \leqslant \alpha \leqslant 1$,

$$f(\alpha) = 2\alpha \frac{4-\alpha^2}{8-\alpha} + \alpha^9 (0.0073 + 0.0621(\alpha - 0.873)^2),$$

to a maximum error of 2×10^{-5}

(b) for $\alpha > 1$,

use $f(\alpha) = \alpha f(1/\alpha)$.

THE RANDOM ORDER SERVICE G/M/m QUEUE

Stig I. Rosenlund

University of Göteborg
Göteborg, Sweden

ABSTRACT

The waiting time in the random order service $G/M/m$ queue is studied. For the Laplace transform we obtain a simpler representation than previously available. For the moments, an explicit recursive algorithm is given and carried out numerically for some cases. This gives rise to the conjecture that the waiting-time distribution can be approximated by the one for $M/M/m$ after a suitable change of scale.

1. INTRODUCTION

Consider the random order service $G/M/m$ queue with interarrival time distribution function $F(F(0) = 0)$ and service time density $\mu e^{-\mu t}$. Let G be the distribution function of the waiting time W (in the stationary case) conditional on its being positive, i.e., $G(t) = P(W \leqslant t \mid W > 0)$. We shall study G through the Laplace transform $\hat{G}(s) = \int_0^\infty e^{-st} dG(t)$ and the moments $\mu_n = \int_0^\infty t^n dG(t)$ of G.

The model has been investigated by LeGall [3] and Takács [6], who give the characteristic function and Laplace transform, respectively, of W. Carter and Cooper [1] and Cooper [2] study G directly and give recursive algorithms for its computation. Carter and Cooper [1] mention that their analysis was motivated by a study of the Bell System's No. 101 Electronic Switching System.

By a substitution of function in Takács' basic differential Eq. (27) we are here able to obtain a simpler closed expression for the Laplace transform than that following from Takács' Eq. (23). Also we give a simpler recursive algorithm for the moments than the one indicated by Takács in his Eq. (32)-(36). The algorithm is carried out numerically for some special cases. The study of moments gives rise to some conjectures on approximations for G.

The $G/M/m/N$ model has been studied by Rosenlund ([5] § 10). By relations (20), (24), (27), (28) and (29) the Laplace transform of W can for $N < \infty$ be calculated without need for numerical integration in the usual $D/M/m$ and $E_K/M/m$ cases, so that it might be preferable to approximate the present infinite waiting room with a finite waiting room. Our notation is

$$\beta = m\mu,$$

$$\hat{F}(s) = \int_0^\infty e^{-st} dF(t),$$

$$\psi(s) = \hat{F}(\beta s),$$

$\rho = (\beta \int_0^\infty t\, dF(t))^{-1}$ (assumed < 1),

$\alpha(s) =$ root z with smallest absolute value of the equation $z = \psi(s + 1 - z)$,

$\omega = \alpha(0)$,

$A =$ as given by (11) in Takács [6] (cited by Cooper ([2], p. 186)),

$W_i =$ waiting time of i th customer,

$V_s(x) = \sum_{k=0}^{\infty} E(e^{-s\beta W_i} | i$ th customer finds $m + k$ other customers at arrival x^k,

$M_n(x) = \sum_{k=0}^{\infty} E((\beta W_i)^n | m + k$ customers before i th arrival$)\, x^k$,

$W =$ random variable with the limiting distribution for W_i as $i \to \infty$.

The notation is adapted to obtain functions which are invariant under changes of time scale.

Let P_k be an arriving customer's distribution for the number of other customers in the system in the stationary (long-run) case. From Takács [6] Eq. (9), we quote

$$P_k = A\omega^{k-m}, k \geqslant m.$$

Hence we can derive the relations

(1) $$P(W \leqslant t) = 1 - A(1 - \omega)^{-1} + A(1 - \omega)^{-1} G(t),$$

(2) $$\hat{G}(s) = (1 - \omega)\, V_{s/\beta}(\omega),$$

(3) $$E(e^{-sW}) = 1 - A(1 - \omega)^{-1} + A(1 - \omega)^{-1} \hat{G}(s),$$

(4) $$\mu_n = (1 - \omega)\beta^{-n} M_n(\omega),$$

(5) $$E(W^n) = A(1 - \omega)^{-1} \mu_n.$$

For $m = 1$, it holds that $A = \omega(1 - \omega)$.

2. THE LAPLACE TRANSFORM

From Takács [6] Eq. (27), we get

(6) $$(x - \psi(s + 1 - x))\, V'_s(x) + V_s(x) = (1 - \psi(s + 1 - x))(1 - x)^{-1}(s + 1 - x)^{-1},\ 0 \leqslant x < 1.$$

The relation between Takács' notation and ours is $\hat{F}(s) = \phi(s)$, $\alpha(s) = \gamma(\beta s)$ and $V_s(x) = \Phi(\beta s, x)$. Equation (6) also follows from Eq. (24) in Rosenlund [5], which was derived by different methods and is in a different form than Eq. (26) in Takács [6]. Before its solution we make the substitution

$$U_s(x) = V_s(x) - (s + 1 - x)^{-1}.$$

Then from Eq. (6)

$$U'_s(x) + U_s(x)/(x - \psi(s + 1 - x)) = s(1 - x)^{-1}(s + 1 - x)^{-2}.$$

Now take $s > 0$ real and let I stand for either $[0, \alpha(s))$ or $(\alpha(s), 1)$. With z an arbitrary fixed point in I we have for x in I

$$\frac{d}{dx}\left[U_s(x) \exp\left\{\int_x^z \frac{dt}{\psi(s + 1 - t) - t}\right\}\right] = s(1 - x)^{-1}(s + 1 - x)^{-2} \exp\left\{\int_x^z \frac{dt}{\psi(s + 1 - t) - t}\right\},$$

whence

$$U_s(x)\exp\left\{\int_x^z \frac{dt}{\psi(s + 1 - t) - t}\right\}$$

(7)

$$= \int_z^x s(1 - u)^{-1}(s + 1 - u)^{-2}\exp\left\{\int_u^z \frac{dt}{\psi(s + 1 - t) - t}\right\}du + C_{s,z}.$$

Setting $x = z$ it is seen that the constant of integration $C_{s,z} = U_s(z)$. Let now $x \to \alpha(s)$. Then $\psi(s + 1 - x) - x \sim (\alpha(s) - x)(1 + \psi'(s + 1 - \alpha(s)))$, so that the left side of Eq. (7) tends to 0. Hence the first term of the right side is $-U_s(z)$ for $x = \alpha(s)$. Put $Q(s) = U_s(\omega)/s$. Then

(8) $$Q(s) = \int_{\alpha(s)}^{\omega} (1 - u)^{-1}(s + 1 - u)^{-2} \exp\left\{\int_u^{\omega} \frac{dt}{\psi(s + 1 - t) - t}\right\}du.$$

The exp factor is $\leqslant 1$. A comparison with Eq. (23) in Takács [6] reveals the relative simplicity of Eq. (8). From Eq. (2) we now get an expression for $\hat{G}(s)$. Making substitutions of variable to get real intervals of integration also for complex s we obtain, letting

$$f_s(y) = 1 - \alpha(s) - (\omega - \alpha(s))y,$$

(9)

$$Q(s) = (\omega - \alpha(s)) \int_0^1 f_s(y)^{-1}(s + f_s(y))^{-2} \exp\left\{\int_y^1 \frac{(\omega - \alpha(s))\,dt}{\psi(s + f_s(t)) + f_s(t) - 1}\right\}dy.$$

The resulting expression for $\hat{G}(s)$ holds also for complex s with $Re(s) \geqslant 0$, and we can use Lévy's inversion formula for characteristic functions, which for distribution functions F such that $F(t) = 0$ for $t < 0$ can be written

(10) $$F(t) = \frac{2}{\pi} \int_0^{\infty} \sin(t\,x)x^{-1}\,\mathrm{Re}(\hat{F}(i\,x))\,dx,$$

if $t \geqslant 0$ is a point of continuity for F. The integral is defined at least in the improper Riemann sense. Inverting $\hat{G}(s)$ we note that $(\beta - \beta\omega)/(s + \beta - \beta\omega)$ is the Laplace transform of the exponential distribution with mean $1/(\beta - \beta\omega)$. This is the distribution of waiting time (conditional on its being positive) in the "first come, first served" $G/M/m$ queue. See Eq. (14) in Takács [6]. We can now state

THEOREM 1: With Q given by Eq. (8) or (9) it holds that

$$\hat{G}(s) = (\beta - \beta\omega)/(s + \beta - \beta\omega) + (1 - \omega)(s/\beta)Q(s/\beta)$$

and

$$G(t) = 1 - e^{-(1-\omega)\beta t} - \frac{2}{\pi}(1-\omega)\int_0^\infty \mathrm{Im}(Q(ix))\sin(\beta t\, x)\,dx.$$

From Eq. (7) we get

$$M_1(x) = -\frac{\partial}{\partial s}\left\{V_s(x)\right\}_{s=0} = (1-x)^{-2} - \int_\omega^x (1-u)^{-3} \exp\left\{\int_u^x \frac{dt}{\psi(1-t)-t}\right\}du. \tag{11}$$

This relation can be used for calculating mean wait when the arriving customer's queue length distribution is not the stationary one. Applying a Tauberian result we can obtain $G'(0) = \lim_{s\to\infty} s\hat{G}(s)$ from Theorem 1. By dominated convergence in Eq. (8) we obtain

$$G'(0) = (\beta - \beta\omega)\omega^{-1}\log(1/(1-\omega)). \tag{12}$$

3. THE MOMENTS

Takács [6] indicates by his Eqs. (32)-(36) a method of calculating the moments $E(W^n)$. We shall here develop a simple and explicit recursive algorithm for this purpose. It is easily shown that

$$M_n(x) = (-1)^n \frac{\partial^n}{\partial s^n}\left\{V_s(x)\right\}_{s=0}. \tag{13}$$

Let us differentiate both sides of Eq. (6) n times in s and r times in x, putting $s = 0$ and $x = \omega$. Simplifying the resulting equation by substituting

$$c_0 = 0,$$
$$c_r = (1-\omega)^{r-1}(-1)^r\psi^{(r)}(1-\omega)/r! \text{ for } r \geqslant 1,$$
$$B_{n,r} = (1-\omega)^{n+r+1}M_n^{(r)}(\omega)/(n!r!),$$

we obtain the following formula, which might be considered the most useful result of this note:

$$B_{n,r} = (1 + r - rc_1)^{-1}\left\{\binom{n+r+1}{r} - rc_1B_{n,r} + \sum_{i=1}^{n}\sum_{k=0}^{r}\binom{n-i+r-k}{n-i}c_{n-i+r-k}\left[(k+1)B_{i,k+1} - \binom{i+k+1}{k}\right]\right\}, \quad n \geqslant 1,\ r \geqslant 0. \tag{14}$$

The terms with $B_{n,r}$ on the right side cancel out, and the term with $B_{n,r+1}$ vanishes, since $c_0 = 0$. Hence Eq. (14) is a recursion. In programming, no regard need be given to the term $-rc_1B_{n,r}$ provided all data registers for the B:s are zero initially. To get $B_{1,0}, B_{2,0}, \ldots, B_{p,0}$ we calculate Eq. (14) for $r = 0, \ldots, p-n$ and $n = 1, \ldots, p$. We get successively

$$B_{1,0}\ B_{1,1} \cdots B_{1,p-2}\ B_{1,p-1},$$
$$B_{2,0}\ B_{2,1} \cdots B_{2,p-2},$$
$$\cdots\cdots\cdots\cdots,$$
$$B_{p-1,0}\ B_{p-1,1},$$
$$B_{p,0}.$$

We need only $c_1, \ldots, c_{p-1}$. For $r > 0$ the interest of $B_{n,r}$ is only as a stepping stone on the way to $B_{1,0}, \ldots, B_{p,0}$. From Eq. (4) then

THEOREM 2: The moments of G are $\mu_n = (\beta - \beta\omega)^{-n} n! B_{n,0}$, where $B_{n,0}$ are obtained recursively from Eq. (14).

Note that the factor $(\beta - \beta\omega)^{-n} n!$ is the nth moment of the conditional waiting time distribution for first come, first served queue mentioned in connection with Theorem 1. Hence $B_{n,0}$ has independent interest as a comparison between disciplines of service with respect to moments.

The recursion (14) is well suited for numerical computation (see Table 1) but to throw more light on the mathematical form of μ_n we go further. Define

$$a_{1,0} = 1,$$
$$e_j = 1 + j - jc_1,$$
$$a_{n,r} = \prod_{j=1}^{n+r-1} e_j^{\min(n,n+r-j)},$$
$$D_{n,r} = a_{n,r} B_{n,r}.$$

Substitution in Eq. (14) gives

$$D_{n,r} = \binom{n+r+1}{r} e_r^{-1} a_{n,r} - rc_1 e_r^{-1} D_{n,r} + \sum_{i=1}^{n} \sum_{k=0}^{r}$$

(15)

$$\binom{n-i+r-k}{n-i} c_{n-i+r-k} \left[(k+1) e_r^{-1} a_{n,r} a_{i,k+1}^{-1} D_{i,k+1} - \binom{i+k+1}{k} e_r^{-1} a_{n,r} \right].$$

As before, the terms with $D_{n,r}$ and $D_{n,r+1}$ on the right side cancel out and vanish, respectively. For all other terms the coefficient $e_r^{-1} a_{n,r} a_{i,k+1}^{-1}$ can be seen to be a polynomial in $e_1, \ldots, e_{n+r-1}$ and hence in c_1. Thus $D_{n,r}$ is a polynomial in $c_1, \ldots, c_{n+r-1}$.

THEOREM 3: It holds that $\mu_n = (\beta - \beta\omega)^{-n} n! D_{n,0} / \prod_{j=1}^{n-1} (1 + j - jc_1)^{n-j}$, where $D_{n,0}$ is a polynomial in $c_1, \ldots, c_{n-1}$ obtained recursively from Eq. (15).

For the first three moments we obtain

(16)
$$\begin{cases} D_{1,0} = 1 \\ D_{2,0} = 2 \\ D_{3,0} = c_2(6 - c_1) + 12 - 8c_1 + 3c_1^2 - c_1^3. \end{cases}$$

Takács [6] gave the first and the second moment. For $n \geqslant 4$ the closed expressions for $D_{n,0}$ are complicated, and the recursion (14) is preferable for obtaining numerical results.

Let us apply the results to the cases of constant and Erlang-distributed interarrival times. For the deterministic case, where $\hat{F}(s) = e^{-sT}$ and $\rho = 1/\beta T$, we define ω by $\omega = \exp\{(\omega - 1)/\rho\}$, $0 < \omega < 1$. It holds that

$$c_r = \omega(1-\omega)^{-1}(-\log(\omega))^r/r!,\ r \geqslant 1. \tag{17}$$

For the gamma (Erlang) case, where $\hat{F}(s) = (\lambda/(s+\lambda))^K$ $(K > 0)$ and $\rho = \lambda/\beta K$, ω is defined by $\omega = (1+(1-\omega)/\rho K)^{-K}$, $0 < \omega < 1$. Here

$$c_r = \omega(1-\omega)^{-1}(1-\omega^{1/K})^r \binom{K-1+r}{r},\ r \geqslant 1. \tag{18}$$

In particular for the M/M/m queue $(K = 1)$ we have $\omega = \rho = \lambda/\beta$ and $c_r = (1-\omega)^{r-1}\omega (r \geqslant 1)$. It follows that for this case

$$\mu_3 = 12(\beta - \lambda)^{-3}(2+\omega)(2-\omega)^{-2}. \tag{19}$$

Table 1 gives $B_{1,0}$, $B_{2,0}$, ..., $B_{10,0}$ for $\hat{F}(s)$ equal to e^{-sT}, $(\lambda/(s+\lambda))^4$, and $\lambda/(s+\lambda)$, i.e., for the queues $D/M/m$, $E_4/M/m$, and $M/M/m$, and for the traffic intensities $\rho = 0.5$, 0.7, and 0.9. We used the calculator TI 59 and run time was 3.75 h for each case, in all 33.75 h.

TABLE 1 — *Values of* $B_{n,0}$ *for* $1 \leqslant n \leqslant 10$

	$D/M/m$			$E_4/M/m$			$M/M/m$		
ρ	0.5	0.7	0.9	0.5	0.7	0.9	0.5	0.7	0.9
ω	.203188	.466996	.806900	.301931	.552912	.843335	.5	.7	.9
1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.25500	1.50053	1.81251	1.28835	1.51643	1.81482	1.33333	1.53846	1.81818
3	1.97820	3.06019	4.76979	2.07762	3.11494	4.77927	2.22222	3.19527	4.79339
4	3.68889	7.69510	16.2385	3.95696	7.89274	16.2857	4.37037	8.18958	16.3561
5	7.75822	22.5330	67.1263	8.51405	23.3153	67.3971	9.72840	24.5066	67.8014
6	17.8677	74.1797	323.777	20.1019	77.5100	325.506	23.8214	82.6458	328.092
7	44.2026	268.155	1773.23	51.0705	283.195	1785.31	62.9102	306.694	1803.40
8	115.889	1046.83	10810.9	137.706	1118.21	10902.1	176.680	1231.30	11038.9
9	318.837	4359.09	72282.5	390.125	4712.69	73019.7	522.226	5281.22	74127.5
10	913.672	19180.3	523851	1152.41	20999.0	530185	1611.84	23968.8	539726

The table illustrates the heavy tail of G in comparison with that of the exponential distribution with mean $1/(\beta - \beta\omega)$, particularly under heavy traffic.

Holding F fixed up to a scale factor and letting $\rho \to 1$ we have $\lim c_1 = 1$ and $\lim c_r = 0$ for $r \neq 1$. This gives in Eq. (14)

$$\lim_{\rho \to 1} B_{n,r} = (n+r)!/r!, \tag{20}$$

so that by Theorem 2

$$\lim_{\rho \to 1} (\beta - \beta\omega)^n \mu_n = (n!)^2. \tag{21}$$

Now $(n!)^2 = E((XY)^n)$, where X and Y are independent with density e^{-x}. We cannot, however, deduce that $\lim G(t/(\beta - \beta\omega)) = P(XY \leqslant t) = \int_0^\infty e^{-y}(1-e^{-t/y})\,dy$ since the moment

sequence $(n!)^2$ does not determine the corresponding distribution uniquely. Neither does it seem possible to establish such a convergence from the Laplace transform $\hat{G}$.

Even for moderately heavy traffic Table 1 reveals that we can write approximately $B_{n,0} \approx (n!)^{\alpha}$ for some $\alpha (0 < \alpha < 1)$ determined by the parameters, so that

$$\mu_n \approx (\beta - \beta\omega)^{-n}(n!)^{1+\alpha}. \tag{22}$$

We can determine α to make Eq. (22) exact for $n = 3$, i.e.,

$$\alpha = \log(B_{3,0})/\log(6), \tag{23}$$

where, by Eq. (16) and Theorem 3,

$$B_{3,0} = [c_2(6 - c_1) + 12 - 8c_1 + 3c_1^2 - c_1^3](2 - c_1)^{-2}(3 - 2c_1)^{-1}.$$

The approximation is not so good for light traffic.

It is further seen from Table 1 that $B_{n,0}$ depends heavily on the traffic intensity ρ but not much, given ρ, on the form of the interarrival distribution F, although μ_n depends strongly on F through ω in the factor $(\beta - \beta\omega)^{-n}$. This suggests that G can be approximated by the distribution for $M/M/m$ after a change of scale. More precisely, let $G_{\lambda,\beta}$ denote G when $F(t) = 1 - e^{-\lambda t}$; then our results would indicate the approximation

$$G(t) \approx G_{\rho,1}(t(\beta - \beta\omega)/(1 - \rho)) \tag{24}$$

for ρ not too small. The indeterminancy of the moment problem still remains, though.

4. THE DISTRIBUTION FUNCTION G FOR M/M/m AND D/M/m

For the queues $M/M/m$ and $D/M/m$, special formulas give more useful results for the calculation of G than the general inversion formula of Theorem 1. For $M/M/m$ the result of Pollaczek [4] seems to be the most convenient. It can be written

$$G_{\rho,1}(t) = 1 - 2(1 - \rho)$$

$$\int_0^{\pi} \frac{\exp\left\{\left\{x + 2\arctan\left(\frac{\sqrt{\rho}\sin x}{1 - \sqrt{\rho}\cos x}\right)\right\}\cot x - t(1 + \rho - 2\sqrt{\rho}\cos x)\right\}\sin x\, dx}{(1 + \rho - 2\sqrt{\rho}\cos x)^2(1 + e^{\pi \cot x})}. \tag{25}$$

For $D/M/m$ the most convenient algorithm seems to be the so-called additional conditioning variable method due to P. J. Burke, described in Cooper [2], pp 229-230. In this case the conditioning variable, the number of arriving customers in $(0,t)$, is deterministic. The algorithm is a recursive scheme, which for $D/M/m$ can be reformulated in the following way. Let

$$H_{j,k}(t) = P(\beta W_i > k/\rho + t \mid m + j \text{ customers before } i\text{th arrival}),$$

$$k = 0, 1, \ldots; 0 \leqslant t \leqslant \rho^{-1}. \tag{26}$$

Then

$$G(t) = 1 - (1 - \omega)\sum_{j=0}^{\infty} \omega^j H_{j,[\beta\rho t]}(\beta t - [\beta\rho t]/\rho),\ t \geqslant 0, \tag{27}$$

and $H_{j,k}$ is determined by the recursion

$$(28)\qquad \begin{cases} H_{j,0}(t) = \sum_{r=1}^{j+1} \dfrac{r}{j+1} \dfrac{t^{j+1-r}}{(j+1-r)!} e^{-t}, \; j \geqslant 0 \\ H_{j,k}(t) = \sum_{r=1}^{j+1} \dfrac{r}{j+1} \dfrac{(1/\rho)^{j+1-r}}{(j+1-r)!} e^{-1/\rho} H_{r,k-1}(t), \; k \geqslant 1; j \geqslant 0. \end{cases}$$

Forming the power series

$$\overline{H}_{k,t}(x) = \sum_{j=0}^{\infty} x^j H_{j,k}(t), \; 0 \leqslant x < 1,$$

we have

$$(29)\qquad G(t) = 1-(1-\omega)\overline{H}_{[\beta\rho t],\beta t - [\beta\rho t]/\rho}(\omega).$$

From Eqs. (28) we obtain the recursion

$$(30)\qquad \begin{cases} \overline{H}_{0,t}(x) = x^{-1} \int_0^x e^{t(u-1)}(1-u)^{-2}du \\ \overline{H}_{k,t}(x) = x^{-1} \int_0^x e^{(u-1)/\rho} \overline{H}'_{k-1,t}(u)\,du, \; k \geqslant 1. \end{cases}$$

This results in

$$(31)\qquad G(t) = 1 - (1-\omega)\omega^{-1} \int_{1-\omega}^{1} e^{-\beta t u} u^{-2} du, 0 \leqslant t \leqslant T.$$

The formula will hold for any arrival distribution F such that $F(T-0) = 0$.

For numerical computations it appears that Eq. (31), when applicable, is better than Eqs. (27) and (28), while for $t > T(\beta t > \rho^{-1})$ generally Eqs. (27) and (28) are better than Eqs. (29) and (30). In Table 2 we study the suggested approximation (24). For each ρ-value, the

TABLE 2 — *Values of G for D/M/m with approximation (24)*

	$\rho = 0.5$		$\rho = 0.8$	
βt	$G(t)$	appr.	$G(t)$	appr.
0	0	0	0	0
0.25	0.1995	0.2315	0.1353	0.1580
0.50	0.3591	0.3962	0.2510	0.2744
0.75	0.4868	0.5169	0.3501	0.3646
1.0	0.5889	0.6077	0.4352	0.4369
1.25	0.6707	0.6775	0.5084	0.4964
1.50	0.7371	0.7322	0.5483	0.5464
1.75	0.7885	0.7756	0.5872	0.5889
2.0	0.8305	0.8106	0.6244	0.6257
2.5	0.8683	0.8624	0.6921	0.6858
3.0	0.9013	0.8981	0.7331	0.7328
4.0	0.9482	0.9413	0.8028	0.8011
5.0	0.9670	0.9647	0.8503	0.8477
7.0	0.9872	0.9859	0.9060	0.9054
10.0	0.9965	0.9958	0.9499	0.9491
17.0	0.9997	0.9996	0.9848	0.9846

left column gives $G(t)$ for $D/M/m$ while the right column gives $G_{\rho,1}(\beta t(1-\omega)/(1-\rho))$. At least for the larger values of the argument the agreement is seen to be good for both traffic intensities. Since $D/M/m$ might be denoted $E_\infty/M/m$ and since the agreement in moments was shown to be better between $E_4/M/m$ and $M/M/m$ than between $D/M/m$ and $M/M/m$, the approximation (24) should be still better for $E_K/M/m$.

REFERENCES

[1] Carter, G.M. and R.B. Cooper, "Queues with Service in Random Order," Operations Research *20*, pp. 389-405 (1972).

[2] Cooper, R.B., *Introduction to Queueing Theory*, Macmillan, New York, (1972).

[3] LeGall, P., Les Systèmes avec ou sans Attente et les Processus Stochastiques. Tome 1. Dunod, Paris (1962).

[4] Pollaczek, F., "La Loi d'attente des appels téléphoniques" Comptes Rendus Académie des Sciences, Paris, *222*, pp. 353-355 (1946).

[5] Rosenlund, S.I., "The Queue G/M/m/N: Busy Period and Cycle, Waiting Time Under Reverse and Random Order Service," Naval Research Logistics Quarterly *25*, pp. 107-119 (1978).

[6] Takács, L., "Delay Distributions for Simple Trunk Groups with Recurrent Input and Exponential Service Times," Bell System Technical Journal *41*, pp. 311-320 (1962).

SOME RESULTS OF THE QUEUEING SYSTEM $E_k^X/M/c$*

D. F. Holman

Canadian Armed Forces

W. K. Grassmann

University of Saskatchewan
Saskatoon, Saskatchewan

M. L. Chaudhry

Royal Military College
Kingston, Ontario

ABSTRACT

This paper analyses the $E_k^X/M/c$ queueing system and shows how to calculate the expected number in the system, both at a random epoch and immediately preceding an arrival. These expectations are expressed in terms of certain initial probabilities which are determined by linear equations. The advantages and disadvantages of this method are also discussed.

1. INTRODUCTION AND BACKGROUND

This paper investigates a c-server queue with exponential service times in which arrivals occur in batches or groups of size X, where X is a random variable. The time between the arrivals of two groups is Erlang-distributed.

The distribution of X, the size of the arriving groups, is denoted by $\{a_i\}$, where $a_i = P(X = i)$. The expected batch size is denoted by $\bar{a}$, which is equivalent to $\sum_{i=1}^{\infty} ia_i$.

The arrival rate of the groups is denoted by λ. Since the time between the two arrivals is Erlang-k distributed, the phases change at a rate of $k\lambda$. Units entering the system join a single queue of unlimited size. The service facility is a group of c identical, exponential channels in parallel, each with a service rate μ. The transition into service is instantaneous in that, when a service channel is free, a unit from the queue (if the queue is not empty) goes into service without delay.

Multiserver queues with batch arrivals have been investigated by a number of authors. In particular, Neuts [8] has recently shown that in the $GI^X/M/c$ queueing system, the queue length distribution is matrix geometric, provided the group size cannot exceed a certain limit g.

*The research for this paper was supported in part by the Defense Research Board of Canada, Grant No. 3610-603, and by Natural Science and Engineering Research Council Canada, Grant No. A8112.

In a more restricted scope, Abol'nikov [1] and Kabak [7] have studied the $M^X/M/c$ queue. The results of these authors were later consolidated and extended by Cromie, et al. [3]. Cromie [3] also corrected the waiting-time distributions of the earlier authors, following a paper of Burke [2].

This paper concentrates on finding the expected number of elements in the system. However, an extended version of this paper is available [5,6] which discusses not only the distribution of the number of elements in the system but also the waiting-time distribution and the related measure of efficiency. The results concerning the mean waiting time W are not being presented here partly because of space constraints and partly because once L is known, W can be evaluated using the Little's formula, $L = \lambda \bar{a} W$.

2. THE EQUILIBRIUM DISTRIBUTION OF QUEUE LENGTH

This section gives equations to determine $p_{n,r}$, the joint equilibrium distribution of n and r, the number of elements in the system with the arrival group being in the rth phase, $1 \leqslant r \leqslant k$. It also derives a generating function for the number of elements in the system, and gives the expected number of elements at a random point in time. Further results can be found in the extended version of this paper [6].

Using the normal approach, the equilibrium equations for the system in question can be found to be

$$(1) \qquad (k\lambda + n\mu)p_{n,r} = (n+1)\mu p_{n+1,r} + k\lambda p_{n,r-1}, \quad \begin{matrix} 0 \leqslant n < c \\ 2 \leqslant r \leqslant k, \end{matrix}$$

$$(2) \qquad (k\lambda + n\mu)p_{n,1} = (n+1)\mu p_{n+1,1} + k\lambda \sum_{i=1}^{n} a_i p_{n-i,k}, \quad n < c,$$

$$(3) \qquad (k\lambda + c\mu)p_{n,r} = c\mu p_{n+1,r} + k\lambda p_{n,r-1}, \quad \begin{matrix} n \geqslant c \\ 2 \leqslant r \leqslant k, \end{matrix}$$

$$(4) \qquad (k\lambda + c\mu)p_{n,1} = c\mu p_{n+1,1} + k\lambda \sum_{i=1}^{n} a_i p_{n-i,k}, \quad n \geqslant c.$$

We also define p_n as the number of elements in the system, that is,

$$p_n = \sum_{r=1}^{k} p_{n,r}.$$

The sum of all p_n must equal one. This can only be accomplished if $\rho < 1$, where ρ is defined as $\bar{a}\lambda/(c\mu)$.

It can be shown that the probability that there are n elements in the system, given that the arrival process is in phase r, is equal to

$$P(n|r) = k\, p_{n,r}.$$

The generating function of $P(n|r)$ will turn out to be important later. We therefore define

$$P_r(z) = \sum_{n=0}^{\infty} P(n|r) z^n = k \sum_{n=0}^{\infty} p_{n,r}\, z^n.$$

Also of importance is $A(z)$, the generating function of the distribution of the arrival group size, which is

$$A(z) = \sum_{n=1}^{\infty} z^n a_n.$$

It is shown in the extended version of this paper that $p_k(z)$ is equal to

$$P_k(z) = \frac{\frac{\mu}{\lambda}(z-1)\sum_{r=1}^{k}\left\{\left[\left(1+\frac{c\mu}{k\lambda}\right)z - \frac{c\mu}{k\lambda}\right]^{r-1} z^{k-r}\sum_{n=0}^{c-1}(c-n)p_{n,r}\,z^n\right\}}{\left[\left(1+\frac{c\mu}{k\lambda}\right)z - \frac{c\mu}{k\lambda}\right]^{k} - A(z)z^k}. \tag{5}$$

If one defines

$$y = \left(1+\frac{c\mu}{k\lambda}\right) - \frac{c\mu}{k\lambda}\,\frac{1}{z},$$

$P_k(z)$ can be transformed to give the following expression:

$$P_k(z) = \frac{\frac{\mu}{\lambda}\left(1-\frac{1}{z}\right)\sum_{r=1}^{k}\sum_{n=0}^{c-1} y^{r-1} z^n (c-n) p_{n,r}}{\left[\left(1+\frac{c\mu}{k\lambda}\right) - \frac{c\mu}{k\lambda}\,\frac{1}{z}\right]^{k} - A(z)}.$$

We now define

$$V(z) = \left[\left(1+\frac{c\mu}{k\lambda}\right) - \frac{c\mu}{k\lambda}\,\frac{1}{z}\right]^{k} - A(z),$$

$$U(z) = \sum_{r=1}^{k}\sum_{n=0}^{c-1} y^{r-1}\, z^n\, (c-n)\, p_{n,r}.$$

It can now be shown that $V(z)$ has $k-1$ zeros inside the unit circile, which we number $z_1, z_2, \ldots z_{k-1}$. For these zeros, $U(z)$ must be zero, which gives $k-1$ equations to determine the $p_{n,r}$

$$U(z_i) = \sum_{r=1}^{k}\sum_{n=0}^{c-1} y_i^{r-1}\, z_i^n\, (c-n)\, p_{n,r} = 0. \tag{6}$$

Here,

$$y_i = \left(1+\frac{c\mu}{k\lambda}\right) - \frac{c\mu}{k\lambda}\,\frac{1}{z_i}.$$

Furthermore, it can be shown that

$$U(1) = \sum_{r=1}^{k}\sum_{n=0}^{c-1} (c-n)\, p_{n,r} = c(1-\rho). \tag{7}$$

Equations (1), (2), (6), and (7) are sufficient to determine $p_{n,r}$, $n < c$, $r = 1, 2, \ldots k$. Once the $p_{n,r}$ are determined, other measures of performance are obtained easily. For instance, the expected number in the system immediately before an arrival epoch can be shown to be

$$(8)\quad L^{-} = \frac{1}{c(1-\rho)} \left\{ \sum_{n=0}^{c-1} (c-n) \sum_{r=1}^{k} \left[n + (r-1)\frac{c\mu}{k\lambda} \right] p_{n,r} + c\rho - \frac{1}{2}(k-1)\frac{c^2\mu}{k\lambda} + \frac{1}{2}\frac{\lambda}{\mu} A''(1) \right\},$$

where

$$A''(1) = \left. \frac{d^2 A(z)}{dz} \right|_{z=1}.$$

For the expected number in the system at a random epoch, one finds

$$(9)\qquad L = \frac{\lambda}{c\mu} \left[\bar{a}\, L^{-} + \frac{1}{2} A''(1) + \bar{a} \right] + \frac{1}{c} \sum_{n=0}^{c-1} (c-n)\, n\, p_n.$$

Once the $p_{n,r}$ are known for $n < c$, the $p_{n,r}$ can also be calculated for $n \geqslant c$ as shown in [6]. Furthermore, it is possible to obtain waiting-time distributions and the related measures of efficiency.

As an application of the theory, consider the $E_2^X/M/c$ queueing system in which the group size follows a geometric distribution. Then

$$a_i = p(1-p)^{i-1}, \qquad i > 0,\ 0 < p < 1,$$

$$A(z) = pz/[1-(1-p)z] = p\left[\frac{1}{z} - (1-p) \right]^{-1}.$$

$V(z)$ now becomes

$$V(z) = \left[\left(1 + \frac{c\mu}{2\lambda} \right) - \frac{c\mu}{2\lambda} \frac{1}{z} \right]^2 - p\left[\frac{1}{z} - (1-p) \right]^{-1} = 0.$$

If $\frac{1}{z}$ is replaced by x, this results in the following equation for x:

$$\left[\left(1 + \frac{c\mu}{2\lambda} \right) - \frac{c\mu}{2\lambda} x \right]^2 [x - 1 + p] - p = 0.$$

One divides this equation by $(x-1)$ and obtains a quadratic equation with the roots x_1 and x_2. One of the roots, say x_1, is greater than one, and this root gives $z_1 = 1/x_1$. One now calculates

$$y_1 = \left(1 + \frac{c\mu}{2\lambda} \right) - \frac{c\mu}{2\lambda} x_1$$

and gets from Eq. (6) in an explicit form

$$\begin{aligned} U(z_1) = {} & (c-0)\, p_{0,1} + (c-0)\, p_{0,2}\, y_1 + \\ & (c-1)\, p_{1,1}\, z_1 + (c-1)\, p_{1,2}\, y_1\, z_1 + \\ & \ldots\ldots + \\ & p_{c-1,1}\, z_1^{c-1} + p_{c-1,2}\, y_1\, z_1^{c-1}. \end{aligned}$$

A similar equation is obtained from Eq. (7):

$$\begin{aligned} U(1) = c(1-\rho) = {} & (c-0)\, p_{0,1} + (c-0)\, p_{0,2} + \\ & (c-1)\, p_{1,1} + (c-1)\, p_{1,2} + \\ & \ldots\ldots + \\ & p_{c-1,1} + p_{c-1,2}. \end{aligned}$$

Equations (1) and (2) give

$$2\lambda p_{0,1} = \mu p_{1,1}$$

$$2\lambda p_{0,2} = \mu p_{1,2} + 2\lambda p_{0,1}$$

$$\vdots \qquad \vdots$$

$$[2\lambda + (c-2)\mu]\, p_{c-2,1} = (c-1)\mu\, p_{c-1,1} + 2\lambda[a_1 p_{c-3,2} + \ldots + a_{c-2} p_{0,2}]$$

$$[2\lambda + (c-2)\mu]\, p_{c-2,2} = (c-1)\mu\, p_{c-1,2} + 2\lambda\, p_{c-2,1}.$$

This gives $2c$ equations for the $2c$ variables $p_{0,1}$, $p_{0,2}$, $p_{c-1,1}$, $p_{c-1,2}$, and we found that these equations can be solved easily for c up to 10. Once the $p_{n,r}$, $0 \leqslant n \leqslant c$, $1 \leqslant r \leqslant k$, are known, L^- and L are given by Eq. (8) and (9).

3. THE POWER AND LIMITATIONS OF THE METHOD

If k is low, the zeros of $V(z)$ inside the unit circle can be found easily. In particular, if $k = 2$ and the group size distribution is geometric, no problem arises, as was shown in the preceding section. But even if $V(z)$ is a polynomial of high degree or transcendental, the zero of $V(z)$ inside the unit circle can be found readily as long as $k = 2$. For instance, we had no problem finding z_1 for the $E_2^X/M/c$ queue with group sizes that are Poisson distributed. In general, the root z_1 can be found as follows. One first shows that z_1 is in the range

$$a < z_1 < b, \tag{10}$$

where

$$a = \left(1 + \frac{c\mu}{k\lambda}\right)^{-1} = \left(1 + \frac{c\mu}{2\lambda}\right)^{-1},$$

$$b = \left(1 + 2\frac{c\mu}{k\lambda}\right)^{-1} = \left(1 + \frac{c\mu}{\lambda}\right)^{-1}.$$

Equation (10) is correct because $V(a) < 0$ and $V(b) > 0$. Since there is a narrow range for z_1, Newton's method or the rule of the false position can be applied to find z_1.

For higher values of k, one has to use complex arithmetic. In order to see this, the following theorem is useful.

THEOREM: If k is odd, all $k - 1$ zeros of $V(z)$ inside the unit circle are complex. If k is even, $V(z)$ has $k - 1$ complex zeros inside the unit circle, and one real zero. Moreover, the real zero satisfies Eq. (10).

Thus, if $k = 3$, one has to find two complex zeros inside the unit circle, and these zeros are conjugate. If $k = 4$, one has a pair of conjugate complex zeros and one real zero, and so on. To find a pair of conjugate complex zeros is feasible. Hence, $k = 3$ and $k = 4$ will not pose unsurmountable problems even if the group size has a transcendental generating function, as is the case if the group size in Poisson. For higher k, the zeros inside the unit circle can be found conveniently, provided $A(z)$ is a real function. The equation $V(z) = 0$ can then be converted to a polynomial, and there are efficient computer algorithms available to solve such polynomials. Indeed, it has been shown in another connection that it takes only a few seconds on a computer to find the roots of a polynomial of degree 50 [4].

For high values of c and k, the solution of ck linear equations that determine the ck variables $p_{n,r}$, $0 \leqslant n \leqslant c-1$, $1 \leqslant r \leqslant k$, may pose problems. In this case one has to use algorithms that make use of the block-diagonal structures of the system. Further problems arise in the case of double roots. These problems are discussed in the extended version of this paper. They can occur only for $k \geqslant 5$ because of the theorem stated above. Moreover, it can be shown that no double roots will occur if the group size is geometric or constant.

ACKNOWLEDGMENTS

The authors would like to thank a referee for making suggestions which led to the improvements of the original draft.

REFERENCES

[1] Abol'nikov, L.M., "A Multichannel Queueing System with Group Arrival of Demands," Engineering Cybernetics *4*, 39-48 (1967).

[2] Burke, P.J., "Delays in Single-Server Queues with Batch Input," Operations Research *23*, 830-833 (1975).

[3] Cromie, M.V., M.L. Chaudhry and W.K. Grassmann, "Further Results for the Queueing System $M^x/M/c$," Journal of the Operational Research Society, to appear.

[4] Grassmann, W.K., "Some Computational Aspects of the $M/E_k/1$ Queue in Steady State," Mimeograph Series 449, Department of Statistics, Purdue University (1976).

[5] Holman, D.F., "Some Problems in the Theory of Bulk Queues," Master Thesis, Department of Mathematics, Royal Military College of Canada, Kingston, Ontario (1977).

[6] Holman, D.F., W.K. Grassman and M.L. Chaudhry, "Some Results of the Queueing System $E_k^X/M/c$," full version available from the authors.

[7] Kabak, I.W., "Blocking and Delays in $M^X/M/c$ Bulk Arrival Queueing Systems," Management Science *17*, 112-115 (1970).

[8] Neuts, M.F., "An Algorithmic Solution to the $GI^x/M/c$ Queue with Group Arrivals," Technical Report No. 7812, Department of Statistics & Computer Science, University of Delaware (1978).

AN APPROXIMATION FOR THE WAITING TIME DISTRIBUTION IN SINGLE SERVER QUEUES

Irwin Greenberg

George Mason University

and

Mathtech Division of Mathematica, Inc.
Arlington, Virginia

ABSTRACT

Single server queues with general interarrival and service times are approximated by queues with two-point (Bernoulli) interarrival times and exponential service times. The parameters are chosen such that the first four moments of the difference of the service times and interarrival times in the approximating system equal those of the original system. The aptness of the approximation is discussed and some examples are presented comparing the exact and approximate waiting time distributions. A more complicated approximation is presented using the dual system (exponential arrivals, Bernoulli service) for those cases where the original approximation cannot be used.

1. INTRODUCTION

The use of the Erlang family of distributions to approximate interarrival and service time distributions in single server queues has been studied by a number of authors. Kotiah, Thompson and Waugh [5] presented an algorithm to solve for the l roots of the functional equation arising in $E_k/E_l/1$ queues. The method depends on l being integer and the solution involves pairs of complex conjugate roots which are used in the expression for the waiting time distribution. Though the algorithm is amenable to hand calculation, the amount of work required to end up with a usable equation involving only real numbers is substantial. More general methods for noninteger l have been suggested (see for example, Wishart [9]), although no simple manual algorithms are given. Marchal and Harris [6] pointed out that the approximation technique can be improved if, instead of fitting distributions separately to the service time $\underline{S}$ and the interarrival time $\underline{T}$, a distribution is fitted to the difference $\underline{U} = \underline{S} - \underline{T}$, since the waiting time distribution depends only on $\underline{U}$ and not on $\underline{S}$ and $\underline{T}$ individually. Their scheme involved finding two Erlang distributions such that the first four moments of the difference of these random variables match the first four moments of the actual difference. Unfortunately, the solution algorithm requires machine calculation.

A simple approach in the spirit of Marchal and Harris is to assume exponential service times and utilize a Bernoulli distribution (B) of interarrival times. With the Bernoulli distribution, the interarrival times take on the value t_1 with probability $1 - p$ and the value $t_2 > t_1$ with probability p. The idea is to approximate a GI/G/1 queue by a B/M/1 queue, where the

parameters p, t_1, and t_2, plus the parameter m of the exponential distribution, are chosen such that the first four moments of $\underline{U}$ in the approximate system match the corresponding moments of the original system. The waiting time distribution will be exponential and the computations are particularly simple. However, as will be seen in the third example further on, this approximation is not always possible. The dual system, the M/B/1 queue, might be used to obtain the waiting time approximation in that case.

The fact that the waiting time distribution can be approximated by an exponential, at least asymptotically, has been known for some time. Kingman [4] derived an exponential approximation to the waiting time distribution for GI/G/1 queues in heavy traffic, and Riordan [7] suggests the exponential approximation to waiting times in M/G/1 queues. As Wishart [9] points out, almost any probability frequency function can be written as a weighted, infinite sum of integer Erlang distributions. A result of Smith [8] can be applied to show that the waiting time distribution can then be written as an infinite sum of negative exponentials. For large values of the random variables, the terms with large (in magnitude) exponents vanish rapidly and the term with the smallest parameter governs the tail of the distribution.

2. THE B/M/1 APPROXIMATION

Let a_i be the ith moment of the interarrival times, and let b_i be the ith moment of the service times. These moments could be either obtained from knowledge of the distribution or calculated from data. The first four moments of $\underline{U}$ are:

$$c_1 = b_1 - a_1$$
$$c_2 = b_2 - 2b_1a_1 + a_2$$
$$c_3 = b_3 - 3b_2a_1 + 3b_1a_2 - a_3$$
$$c_4 = b_4 - 4b_3a_1 + 6b_2a_2 - 4b_1a_3 + a_4.$$

If the B/M/1 queue is used as an approximation, then the ith moment of the approximating interarrival distribution is

$$a_i' = (1 - p)t_1^i + pt_2^i,$$

and the ith moment of the approximating service distribution is

$$b_i' = i!/m^i.$$

The moments of the difference, c_i' are obtained as are the c_i with a_i' and b_i' replacing a_i and b_i respectively. Setting $c_i' = c_i$ for $i = 1, 2, 3, 4$ yields

$$\frac{1}{m} - [(1 - p)t_1 + pt_2] = c_1$$
$$\frac{2}{m^2} - \frac{2}{m}[(1 - p)t_1 + pt_2] + [(1 - p)t_1^2 + pt_2^2] = c_2$$
$$\frac{6}{m^3} - \frac{6}{m^2}[(1 - p)t_1 + pt_2] + \frac{3}{m}[(1 - p)t_1^2 + pt_2^2] - [(1 - p)t_1^3 + pt_2^3] = c_3$$
$$\frac{24}{m^4} - \frac{24}{m^3}[(1 - p)t_1 + pt_2] + \frac{12}{m^2}[(1 - p)t_1^2 + pt_2^2] - \frac{4}{m}[(1 - p)t_1^3 + pt_2^3] + [(1 - p)t_1^4 + pt_2^4] = c_4.$$

These four equations must be solved simultaneously for m, p, t_1, and t_2. The solution is tedious but the final result is surprisingly simple. The parameter m is the solution to the cubic equation

$$m^3 + K_1 m + K_2 = 0, \tag{1}$$

with

$$K_1 = \frac{4c_1c_3 - c_4 - 3c_2^2}{2c_1c_2c_3 + c_2c_4 - c_1^2c_4 - c_3^2 - c_2^3},$$

and

$$K_2 = 4\frac{c_3 - 3c_1c_2 + 2c_1^3}{2c_1c_2c_3 + c_2c_4 - c_1^2c_4 - c_3^2 - c_2^3}.$$

The extraction of the roots of a cubic equation is discussed in most college algebra texts and mathematical handbooks, for example, Burington [1].

The remaining parameters are calculated as follows: let

$$K_3 = \frac{(c_1c_3 - c_2^2)m^4 + (c_1c_2 - c_3)m^3 + (3c_2 - 4c_1^2)m^2}{(c_2 - c_1^2)m^2 - 1}$$

and

$$K_4 = \frac{K_3 + c_2m^2 - 2c_1m}{1 - c_1m}.$$

The larger root of the quadratic equation

$$y^2 - K_4y + K_3 = 0$$

is denoted by y_2 and the smaller root by y_1. Then

$$p = \frac{1 - y_1 - c_1m}{y_2 - y_1} \tag{2}$$

$$t_1 = \frac{y_1}{m}, \qquad t_2 = \frac{y_2}{m}.$$

The cubic equation in m may have more than one real positive solution. The experience of the author in working examples is that when this occurs (and it will occur if the third moment of the service time distribution about its mean exceeds the third moment of the interarrival distribution about its mean) none of the real positive solutions will yield positive values for p, t_1, and t_2. The B/M/1 queue cannot be used as an approximation in that case.

If m, p, t_1, and t_2 are obtained as above, then the equation

$$(1 - p)e^{-m(1-z)t_1} + pe^{-m(1-z)t_2} - z = 0 \tag{3}$$

has one real root in the interval, $0 < z < 1$; call this root Z. The expected waiting time prior to service in this approximate system is

$$W_q' = \frac{Z}{(1 - Z)m}.$$

The probability that an arrival does not have to wait is

$$W_q'(0) = 1 - Z,$$

and the probability of a wait longer than t prior to the start of service is

$$1 - W_q'(t) = Ze^{-m(1-Z)t}$$

(see Gross and Harris [3], pages 277-279).

3. SOME EXAMPLES

EXAMPLE 1: For the $E_k/E_l/1$ queue with $k = 1.5$, $1/\lambda = 2.5$, $l = 6$, $1/\mu = 2$, the algorithm of Kotiah et al. [5] yields the exact probability of a wait greater than t:

$$\begin{aligned} 1 - W_q(t) &= 0.802e^{-0.252t} - 0.013e^{-4.740t} + 0.050e^{-2.271t} \\ &\quad [\cos(1.234 + 1.805t) - 0.649\cos(0.047 + 1.805t)] \\ &\quad + 0.021e^{-4.008t}[\cos(2.193 + 1.470t) - 0.594\cos(0.021 + 1.470t)]. \end{aligned}$$

The approximating system is based on the interarrival moments

$$a_i = \left(\frac{2.5}{1.5}\right)^i \prod_{j=0}^{i-1} (1.5 + j)$$

and service time moments

$$b_i = \left(\frac{2}{6}\right)^i \prod_{j=0}^{i-1} (6 + j).$$

These are used to calculate the c_i and K_1 and K_2. The cubic equation becomes

$$m^3 - 0.5486m - 0.1405 = 0$$

which has $m = 0.8454$ as its single real positive solution. K_3 and K_4 can now be calculated to yield the quadratic equation

$$y^2 - 6.9699y + 5.4373 = 0.$$

The roots are $y_1 = 0.8951$, $y_2 = 6.0748$. Thus, $p = 0.1019$, $t_1 = 1.0587$, $t_2 = 7.1854$, and the approximating distribution is

$$1 - W_q'(t) = 0.711e^{-0.244t}.$$

The exact and approximating distributions are compared in Table 1 as are the means of the interarrival and service time distributions, the mean waiting times, and the traffic intensities. These expectations are shown to demonstrate that in spite of the fact that the approximating system does not resemble the exact system, the waiting time distributions are close to each other.

EXAMPLE 2: For the queue with hyperexponential interarrivals with probability density function

$$a(t) = (2/9)e^{-(2/3)t} + (8/9)e^{-(4/3)t}$$

and hyperexponential service with density function

$$b(t) = (1/4)e^{-t} + (9/4)e^{-3t},$$

the probability of a wait longer than t is

$$1 - W_q(t) = 0.420e^{-0.611t} + 0.106e^{-2.327t}$$

(see Greenberg [2]). The interarrival moments are

$$a_i = i!(3^{i-1} + 1.5^{i-1})/2^i,$$

TABLE 1 — *Comparison of Exact and Approximate B/M/1 Systems*

	Example 1		Example 2	
	Exact	Approximate	Exact	Approximate
Mean interarrival time	2.5	1.683	1	1.377
Mean service time	2	1.183	0.5	0.877
Traffic intensity	0.8	0.703	0.5	0.636
Average delay	3.16	2.92	0.73	0.78
Prob {delay > t}				
t				
0	0.748	0.711	0.526	0.470
1	0.619	0.557	0.238	0.257
2	0.485	0.437	0.125	0.140
3	0.377	0.342	0.067	0.077
4	0.293	0.268	0.036	0.042
5	0.227	0.210	0.020	0.023
6	0.177	0.165	0.011	0.013
7	0.137	0.129	0.006	0.007
8	0.107	0.101	0.003	0.004
9	0.083	0.079	0.002	0.002
10	0.065	0.062	0.001	0.001

and the service time moments are

$$b_i = i!(4^{i-1} + (4/3)^{i-1})/4^i.$$

The cubic equation becomes

$$m^3 - 1.0335m - 0.3057 = 0,$$

which has $m = 1.1408$ as its single real positive solution. The approximating distribution is

$$1 - W_q'(t) = 0.470e^{-0.604t},$$

and the exact and approximate systems are compared in Table 1.

EXAMPLE 3: For the queue with Erlang interarrivals with $k = 3$ and $1/\lambda = 2/3$, and hyperexponential services with density function

$$b(t) = 0.04e^{-0.4t} + 3.24e^{-3.6t},$$

the probability of a wait longer than t is

$$1 - W_q(t) = 0.587e^{-0.168t} + 0.079e^{-2.863t}$$

(see Greenberg [2]). The moments are

$$a_i = \left(\frac{2/3}{3}\right)^i \prod_{j=0}^{i-1} (3 + j)$$

and

$$b_i = i!(10^{i-1} + (10/9)^{i-1})/4^i.$$

The cubic equation becomes

$$m^3 - 2.0457m + 0.7495 = 0,$$

which has two real positive solutions: $m = 1.899$ or 0.3969. Using the first, $K_3 = -15.11$ and using the second $K_3 = -0.11$. Thus the quadratic equation in y will have a single positive root and the appropriate solutions for p, t_1, and t_2 do not exist. In this case, no B/M/1 queue can be used as an approximation.

4. THE M/B/1 APPROXIMATION

If the M/B/1 queue is used as an approximation with L as the parameter of the exponential interarrival time distribution and service times u_1 (with probability $1 - q$) and $u_2 > u_1$ (with probability q), then the ith moment of the approximating interarrival distribution is $a_i' = i!/L^i$ and the ith moment of the approximating service time distribution is $b_i' = (1 - q)u_1^i + qu_2^i$. The ensuing cubic equation which must be solved for L differs from Equation (1) only in one sign:

$$L^3 + K_1 L_1 - K_2 = 0$$

with K_1 and K_2 as in Section 2. Note that the real positive roots of this equation are the absolute values of the real negative roots of Equation (1).

The remainder of the derivation is parallel to that of the B/M/1, resulting in a quadratic equation similar to the one in Section 2:

$$y^2 - K_6 y + K_5 = 0$$

where

$$K_5 = \frac{(c_1c_3 - c_2^2)L^4 - (c_1c_2 - c_3)L^3 + (3c_2 - 4c_1^2)L^2}{(c_2 - c_1^2)L^2 - 1},$$

and

$$K_6 = \frac{K_5 + c_2L^2 + 2c_1L}{1 + c_1L}.$$

The two roots are y_2 and $y_1 < y_2$. The estimate of q is similar to the estimate of p, Equation (2), again differing only in one sign:

$$q = \frac{1 - y_1 + c_1L}{y_2 - y_1}$$

and

$$u_1 = y_1/L, \quad u_2 = y_2/L.$$

The expected waiting time prior to service for this approximating M/B/1 queue is

$$W_q' = \frac{Lb_2'}{2(1 - r)},$$

where r is the traffic intensity of the approximating system given by

$$r = L[(1 - q)u_1 + qu_2].$$

The Laplace-Stieltjes transform of the Bernoulli service time distribution is $(1 - q)\exp(-\theta u_1) + q\exp(-\theta u_2)$. This can be substituted into the Pollaczek-Khinchine formula. This latter expression can be expanded in an infinite series and the binomial theorem invoked twice to yield the transform of the approximating waiting time distribution:

$$\int_{0-}^{\infty} \exp(-\theta t)\, dW_q'(t) =$$

$$(1-r) \sum_{j=0}^{\infty} L^j \sum_{k=0}^{j} \sum_{i=0}^{k} \frac{[-(1-q)]^{k-i}(-q)^i}{(k-i)!\,i!\,(j-k)!} \frac{j!}{\theta^j} \exp\{-\theta[iu_2 + (k-i)u_1]\}.$$

Reversing the order of summation, summing over j, and multiplying by θ yields the transform of $W_q'(t)$, which can be inverted term by term and subtracted from unity to yield the probability of a wait longer than t:

$$1 - W_q'(t) = 1 - (1-r) \sum_{k=0}^{I(1)} (-1)^k \sum_{i=0}^{I(2)} \frac{\{qL[t - iu_2 - (k-i)u_1]\}^i}{i!}$$

$$\cdot \frac{\{(1-q)L[t - iu_2 - (k-i)u_1]\}^{k-i}}{(k-i)!} \exp\{L[t - iu_2 - (k-i)u_1]\}$$

where $I(1)$ is the largest integer less than or equal to t/u_1 and $I(2)$ is the largest integer less than or equal to $(t - ku_1)/(u_2 - u_1)$.

This expression, although a finite sum, is quite difficult to evaluate numerically even using double precision arithmetic on a computer. The approximation to the queue of example 3 requires a sum of thirty terms for the probability of a wait longer than ten units, a probability of about 0.1, with a number of the terms exceeding 10^9 in magnitude.

An "approximation to the approximation" can be used to obtain a simpler result, using the Riordan approximation alluded to earlier. For the M/B/1 queue the Riordan approximation becomes

$$1 - W_q'(t) \approx \frac{1 - L[(1-q)u_1 + qu_2]}{L[(1-q)u_1 e^{-L(1-Z)u_1} + qu_2 e^{-L(1-Z)u_2}] - 1} \exp\{L(1-Z)t\}$$

where $Z > 1$ is the largest root of the analog of Equation (3):

$$(1-q)e^{-L(1-z)u_1} + qe^{-L(1-z)u_2} - z = 0.$$

EXAMPLE 3 (*continued*): The cubic equation

$$L^3 - 2.0457L - 0.7495 = 0$$

has one real positive solution, $L = 1.5868$. Note that $m = -1.5868$ is the third root of the cubic equation for m in the first part of this example. K_5 and K_6 can now be calculated to yield the quadratic equation

$$y^2 - 15.9005y + 8.9137 = 0$$

with roots $y_1 = 0.5819$, $y_2 = 15.3187$. Hence $q = 0.0104$, $u_1 = 0.3667$, and $u_2 = 9.6538$. The approximating traffic intensity is $r = 0.735$.

The Pollaczek-Khinchine sum for $1 - W_q'(t)$ can be evaluated at this point. The Riordan approximation requires the largest root of

$$(1 - 0.0104)e^{-(1.5868)(1-z)(0.3667)} + 0.0104e^{-(1.5868)(1-z)(9.6538)} - z = 0$$

which is $z = 1.1073 \equiv Z$. Hence,

$$1 - W_q'(t) \approx 0.6057e^{-0.1703t}.$$

The exact probability of having to wait longer than time t is contrasted with the M/B/1 approximation based on the Pollaczek-Khinchine formula and with the Riordan approximation in Table 2. The expected waiting time of the approximating M/B/1 system is $W_q' = 3.3$ compared to the true value for the original system of 3.5.

TABLE 2 — *Approximations for Example 3*

t	Probability of Delay > t		
	Exact	Approximations	
		M/B/1	Riordan
0	0.666	0.736	0.606
0.5	0.559	0.483	0.555
1	0.501	0.389	0.511
2	0.420	0.339	0.431
3	0.355	0.312	0.363
4	0.300	0.285	0.306
5	0.253	0.257	0.258
6	0.214	0.227	0.218
7	0.181	0.197	0.184
8	0.153	0.166	0.155
9	0.129	0.133	0.131
10	0.109	0.107	0.110

REFERENCES

[1] R.S. Burington, *Handbook of Mathematical Tables and Formulas* (Handbook Publishers, Inc., Sandusky, Ohio, pp. 7-9, 1953).

[2] I. Greenberg, "Single Server Queues with Hyperexponential Service Times," Naval Research Logistics Quarterly, *24*, pp. 451-455 (1977).

[3] D. Gross and C.M. Harris, *Fundamentals of Queueing Theory*, (Wiley, New York, 1974).

[4] J.F.C. Kingman, "The Heavy Traffic Approximation in the Theory of Queues," in *Proceedings of the Symposium on Congestion Theory* (University of North Carolina Press, Chapel Hill, North Carolina, 1962).

[5] T.C.T. Kotiah, J.W. Thompson and W.A.O. Waugh, "Use of Erlangian Distributions for Single Server Queueing Systems," Journal of Applied Probability, *6*, pp. 584-593 (1969).

[6] W.G. Marchal and C.M. Harris, "A Modified Erlang Approach to Approximating GI/G/1 Queues," Journal of Applied Probability, *13*, pp. 118-126 (1976).

[7] J. Riordan, *Stochastic Service Systems* (Wiley, New York, New York, 1962).

[8] W.L. Smith, "On the Distribution of Queueing Times," Cambridge Philosophical Society Proceedings, *49*, pp. 449-461 (1953).

[9] D.M.G. Wishart, "A Queuing System with Service Time Distributions of Mixed Chi-Squared Type," Operations Research, *7*, pp. 174-179 (1959).

CONGESTION TOLLS: EQUILIBRIUM AND OPTIMALITY

Robert W. Rosenthal

Bell Telephone Laboratories, Inc.
Murray Hill, New Jersey 07974

ABSTRACT

An example of a network with flow costs depending on congestion is presented for which no system of tolls and subsidies exists which can ensure that all equilibria in the game of route selection are Pareto optimal.

It has long been recognized that in transportation facilities subject to congestion delays, equilibrium flows may not be Pareto optimal. The imposition of tolls has been a much discussed remedy (see, for example, the classical treatments by Pigou [4] and Knight [3] and the more recent discussions in Kahn [2] and Edelson [1]). The purpose of this note is to show that an all-knowing central authority, using a system of tolls or a system of tolls and subsidies, cannot in general force an outcome which is Pareto optimal or even Pareto superior to a given nonoptimal equilibrium.

Consider a network containing only two nodes, A and B, which are connected by two distinct arcs, 1 and 2. Four players, 1, 2, 3, and 4, must travel from A to B. The cost (or travel time) incurred by player i on arc k when the total traffic on arc k is x_k is

$$c_{ik}(x_k) = \left\{\begin{array}{ll} 2 & \text{if } i \neq k,\ x_k \leqslant 2 \\ 4 & \text{if } i \neq k,\ x_k > 2 \\ 3 & \text{if } i = k,\ x_k \leqslant 2 \\ 5 & \text{if } i = k,\ x_k > 2 \end{array}\right\} \text{for } i = 1, 2$$

$$= 0 \text{ otherwise.}$$

Thus, players 1 and 2 incur uniformly higher costs on arcs 1 and 2, respectively, while the costs for players 3 and 4 are 0, independent of route and traffic. It is easy to see that the Pareto optimal (costs being minimized) route selections result in the unique cost vector (2,2,0,0) and involve player 1 travelling on arc 2, player 2 travelling on arc 1, player 3 travelling on either arc, and player 4 travelling on the arc not traversed by 3. The set of Nash equilibria of the pure strategy game of route selection, however, consists of all the configurations of two players on each arc. The resulting cost vectors are (2,2,0,0), (2,3,0,0), (3,2,0,0) and (3,3,0,0).

Suppose that a central authority, knowing the above data and fearing a nonoptimal equilibrium, wishes to force a Pareto optimal selection of routes for the players. Consider first a system of tolls on the arcs in the form of some otherwise worthless medium (say tokens) together with subsidies of tokens to the players. (In our usage a toll is a fixed charge which must be paid by every unit of flow using a particular arc. A subsidy is a gift to a player which is independent of route selection. Different players may receive different subsidies, and different arcs may be assigned different tolls; but different players pay the same toll if they use the same arc.) No matter what the subsidies are (as long as all players can afford at least one of the arcs) there is no way to stop players 3 and 4 from travelling on the arc with the cheaper toll (or on the same arc, if both tolls are equal). This cannot result in a Pareto optimal selection.

Similarly, if the tolls and subsidies are in the form of money, 3 and 4 cannot be kept from travelling on the cheaper arc.

A trivial way out of the dilemma in general is to use a different kind of token on each arc in the network and to subsidize each player exactly the required number of units of each kind of token for the route desired. This method, of course, can be used to force every player to use the route deemed desirable for him. It does not appear to be practical in large networks, however.

ACKNOWLEDGMENT

I have benefitted from discussions with Carl Futia and William Taylor on this subject.

REFERENCES

[1] Edelson, N.M., "Congestion Tolls Under Monopoly," American Economic Review *61*, 873-882 (1971).

[2] Kahn, A.E., *The Economics of Regulation: Principles and Institutions,* Vol. I, Chapter 4, Wiley, New York (1970).

[3] Knight, F.H. "Some Fallacies in the Interpretation of Social Cost," Quarterly Journal of Economics *38,* 582-606 (1924). Reprinted in G.J. Stigler and K. Boulding (eds.) *Readings in Price Theory,* Irwin, Chicago, pp. 160-179 (1952).

[4] Pigou, A.C., *The Economics of Welfare,* Macmillan, London, 1920 edition only.

COMPUTING EQUILIBRIA VIA NONCONVEX PROGRAMMING*

Jonathan F. Bard

University of Massachusetts—Boston
Boston, Massachusetts

James E. Falk

School of Engineering and Applied Science
The George Washington University
Washington, D.C.

ABSTRACT

The problem of determining a vector that places a system in a state of equilibrium is studied with the aid of mathematical programming. The approach derives from the logical equivalence between the general equilibrium problem and the complementarity problem, the latter being explicitly concerned with finding a point in the set $S = \{x : < x, g(x) > = 0, g(x) \geqq 0, x \geqq 0\}$. An associated nonconvex program, $\min\{- < x, g(x) > : g(x) \leqq 0, x \geqq 0\}$, is proposed whose solution set coincides with S. When the excess demand function $g(x)$ meets certain separability conditions, equilibrium solutions are obtained by using an established branch and bound algorithm. Because the best upper bound is known at the outset, an independent check for convergence can be made at each iteration of the algorithm, thereby greatly increasing its efficiency. A number of examples drawn from economic and network theory are presented in order to demonstrate the computational aspects of the approach. The results appear promising for a wide range of problem sizes and types, with solutions occurring in a relatively small number of iterations.

1. INTRODUCTION

In this paper we investigate a procedure for computing equilibria from the vantage point of mathematical programming. A competitive model of an ecomony will serve as the basis for the discussion although a variety of contexts would have been equally suitable. Other types of equilibrium problems, such as those arising in traffic network analysis, have direct conceptual and analytic counterparts to those found in economics, and are hence amenable to the same solution techniques.

A state of equilibrium exists when competing or opposing forces are brought into balance. One of the major themes of ecomonic theory is that the behavior of a complex economic system can be viewed as an equilibrium arising from the interaction of a number of economic units, each motivated by their own special interest. General equilibrium theory ([2],[22],[24]) seeks to determine the point at which this balance can be struck, and in so doing focuses on the interrelationships that exist among the markets for goods and services in the economy. The analysis, however, is carried out in terms of individual decision makers and commodities rather than in terms of aggregates. The fundamental questions that general equilibrium theory

*Research sponsored by the U.S. Army Research Office, Durham, N.C.

attempts to answer are the same as those posed in macroeconomic theory: given different economic environments, what goods will the economy produce, how will these be produced, and who will obtain them? But where macroeconomics provides answers in terms of aggregates, general equilibrium theory provides answers in terms of the individual consumers, producers, and commodities making up these aggregates.

Consider, for the moment, a model in which m consumers are engaged in the exchange of commodities which they initially own and in which production or supply is ignored. Suppose there are n goods in the economy and that each of the consumer's preferences is represented by a utility function. A bundle of goods x is preferred to a bundle x' by consumer i $(i = 1, \ldots, m)$ if and only if $u_i(x) > u_i(x')$ where the utility function $u_i : R^n \to R$ is generally assumed to be strictly concave and continuous. Let $p \in R^n$ be the vector of prices for the n goods. The demands of the i th consumer are determined by the solution to the following problem:

$$\begin{aligned} &\text{maximize} && u_i(x) \\ &\text{subject to} && \langle p,x\rangle \leqq \langle p,w^i\rangle \\ & && x \geqq 0\,, \end{aligned}$$

where $w^i \in R^n$ is the initial wealth or resource endowment of the i th consumer, $i = 1, \ldots, m$. We shall assume that the solution vector for this problem, $d^i(p)$, can be written as a continuous function of the prices p. The individual trader's excess demand function is $d^i(p) - w^i$ $(i = 1, \ldots, m)$ and will be denoted by $g^i(p)$. The excess demand will be positive for those commodities whose stock he wishes to increase by exchange and negative for the remaining items. If it is assumed that all purchases are to be financed solely by the sale of assets, then individual budgetary constraints lead to the following identity:

$$p_1 d_1^i(p) + \cdots + p_n d_n^i(p) = p_1 w_1^i + \cdots + p_n w_n^i. \tag{1}$$

The market excess demand function $g : R^n \to R^n$ is simply the sum of the individual excess demand functions

$$g(p) = \sum_{i=1}^{m} (d^i(p) - w^i).$$

An equilibrium price vector p^* is one for which all of the market excess demands are less than or equal to zero with a zero price for any commodity whose excess demand is strictly less than zero. This leads to the formulation of the complementarity problem ([8],[17]):

$$g(p) \leqq 0, \tag{2}$$

$$p \geqq 0, \tag{3}$$

$$\langle p,g(p)\rangle = 0, \tag{4}$$

whose solution p^* will be the focal point of this paper. Condition (4), known as the Walras Law [27], is the aggregated form of Eq. (1) and holds for all price vectors p whether they are in equilibrium or not.

We note here that production may easily be incorporated in this model by either replacing or augmenting the i th consumer's initial wealth w^i by a supply function. For individual i, this function relates the prevailing market prices p to the quantity of goods produced.

A number of persons, including Nash [20], Arrow and Debreu [1], and Kuhn [16] have studied the existence problem of the competitive model from the standpoint of combinatorial

topology. The first algorithms, however, actually designed for computing economic equilibria were developed by Scarf [23], and were based on a procedure for approximating a fixed point of a continuous mapping. More recently, Wilson [28] and Elken [10] have exploited path methods in the pursuit of greater computational efficiency. In a slightly different vein, Lemke [17] offered some constructive proofs relating to the existence of equilibrium points for bimatrix games. His work strongly suggested a computational scheme for models with linear excess demand functions.

This paper presents an alternative procedure for computing equilibria for a class of problems where the excess demand function or its logical equivalent has an explicit representation that can be converted to a separable form. Solutions are obtained by first recasting the complementarity problem into a nonconvex minimization problem whose optimal value or best upper bound is known at the outset, and then using Falk's [12] algorithm to locate a global solution. This allows us to go beyond the common linear formulations of an economy or network (e.g., see Eaves [8], Negishi [21], or Asmuth, Eaves, and Peterson [4]) which, in spite of their outward simplicity, must appeal to rather complicated algorithms if solutions are to be obtained.

The algorithm which we subsequently describe and use as an alternative for solving Eqs. (2-4) is based on a branch and bound philosophy, and as such, computes a convergent sequence of upper and lower bounds on the optimal value of the problem. In our case, however, because the best upper bound on the objective function is known to be zero, the amount of work necessary to achieve convergence is significantly reduced. The usual requirement of finding a point that yields equality between the best upper and best lower bounds is replaced by the simpler requirement of finding any point that yields an objective value of zero.

In the next section, the complementarity problem is reformulated as a nonconvex minimization problem whose solution yields the desired equilibrium vector. Next, the method is applied to a number of sample problems and our computational experience is detailed. Here we see that the results are obtained in a surprisingly small number of iterations of the algorithm.

2.0 REFORMULATION OF THE COMPLEMENTARITY PROBLEM

In the complementarity problem derived above, there is no objective function to be optimized. Indeed, in many complex economic equilibrium problems there does not appear to be a "natural" objective function whose optimization yields prices and quantities in equilibrium (see, e.g., Scarf [24]).

In spite of this, consider the following "artificial" minimization problem (P):

$$\text{(P)} \qquad v^* = \min\{-\langle p, g(p)\rangle : g(p) \leqq 0,\ p \geqq 0\}.$$

Now let p^* be a solution of the complementarity problem (i.e., p^* is a vector of equilibrium prices). Then p^* is feasible to problem (P), and yields a value of 0 to the objective function. Since this objective function is greater than or equal to zero at all feasible points, $v^* = 0$. Conversely, it is clear that any solution of problem (P) for which $v^* = 0$ must be a vector of equilibrium prices.

Problem (P) is of a nonconvex nature, and in general, no suitable technique exists for the determination of a global, rather than a local solution; however, if each excess demand function g_i, $i = 1, 2, \ldots, n$, is separable, i.e.,

$$g_i(p) = \sum_{j=1}^{n} g_{ij}(p_j), \quad i = 1, 2, \ldots, n$$

and each g_{ij} is continuous, then problem (P) can be written as a separable programming problem whose approximate global solution can be obtained with arbitrary precision.

We now formulate an equivalent problem with a different objective function but the same constraint region whose optimal value is equal to that of problem (P). The equivalent problem (P′) is

$$\text{(P′)} \qquad \min\left\{\sum_i \min(p_i, -g_i(p)) : p \geqq 0,\ g(p) \leqq 0\right\}.$$

Rewriting the objective function in problem (P′), we get the desired result:

$$\begin{aligned}
&\min_{\substack{g(p)\leqq 0\\ p\geqq 0}} \left\{\sum_i (\min(0, -g_i(p) - p_i) + p_i)\right\} \\
&= \min\left\{\sum_{\substack{i\\ g(p)\leqq 0\\ w + p + g(p) = 0\\ p\geqq 0}} (\min(0, w_i) + p_i)\right\} \\
&= \min\left\{\sum_i (\min(0, w_i) + p_i)\right\} \qquad \text{(S)} \\
&\quad \sum_j g_{ij}(p_j) \leqq 0 \\
&\quad w_i + p_i + \sum_{\substack{j\\ p\geqq 0}} g_{ij}(p_j) = 0 \qquad i = 1, 2, \ldots, n
\end{aligned}$$

where the w_i's will be referred to as auxiliary variables.

Problem (S) is still a nonconvex programming problem, but its separable structure, created at the expense of a twofold increase in dimensionality, makes its mathematics much more tractable. The traditional method for treating separable problems involves calculating piecewise linear approximations of the associated functions and applying a modification of the simplex method to the resulting problem (see, e.g., Miller [19]). The modification amounts to a restriction on the usual manner of selecting variables to exchange roles (basic to nonbasic and vice versa) and will yield a local but not necessarily a global solution of the approximating problem.

An algorithm for finding global solutions of nonconvex separable problems was developed by Falk and Soland [13] and Soland [25]. The method is based on the branch and bound philosophy and yields a (generally infinite) sequence of points whose cluster points are global solutions of the problem. The implementation of the method is limited by the necessity of computing convex envelopes [11] of the functions involved, although a number of applications have been shown possible when these functions exhibit special structures (e.g., concave or piecewise linear).

The inherent limitations that special problem structures impose have been overcome by the introduction of two algorithms independently developed by Beale and Tomlin [5] and Falk [12]. For this paper, we have used the programming code MOGG based on the algorithm proposed by Falk and written by Grotte [15] to solve a number of equilibrium problems. The results are presented in the next section.

3.0 COMPUTATIONAL EXPERIENCE

A variety of equilibrium problems have been studied to test the approach outlined above. The first is a multicommodity, transshipment problem defined on an affine network, taken from Asmuth, Eaves, and Peterson [4] who used Lemke's algorithm [17] to obtain a solution. The second involves a simple competitive market comprising three producers, three consumers, and three commodities. The supply and demand functions in this economy are given a piecewise linear formulation, and three equilibrium points are known to exist. The third problem is identical to the second except that a majority of the piecewise linear functions have been recast as continuous, smooth functions. The fourth problem provides an example outside the context of economics, and is derived from a 3-node, 4-arc traffic network whose equivalent excess demand function is both nonlinear and nonseparable.

The algorithm itself is based on branch and bound techniques and considers subsets of a linear polyhedron containing the feasible region of problem (S). A lower bound on the optimal value of the problem is found by minimizing the objective function over each of these subsets and selecting the smallest value obtained. A check for the solution is made which, if successful, yields a global solution of the piecewise linear approximation to problem (S). If the check fails, the subset corresponding to the smallest lower bound is further subdivided into either two or three new linear polyhedra and the process continues as before with new and sharper bounds being determined. The process is finite and terminates with a global solution of the approximate problem.

3.1 Transshipment on Affine Networks

Economic equilibria on certain affine, multicommodity, transshipment networks were first studied by the regional economists Takayama and Judge [26] using quadratic programming. Recently, Asmuth, Eaves, and Peterson [4] have constructed a more general approach that utilizes the economic equilibrium conditions directly without first passing to a quadratic programming problem. A brief discussion of their model and the solution to the sample problem presented in their paper follow.

The transshipment problem can conveniently be represented by a directed graph (N,L) with a finite number of nodes (members of N) and links (members of L) on which a finite number of commodities can be transported. Each node i in N represents the set of producers and/or consumers at a specific spatial location; and each link ij in L represents a specific facility for transporting commodities from some node i to a different node j. (In particular, we assume that there are no loops; i.e., no links connecting a given node to itself.) Each link is aligned to coincide with a direction of possible transport; therefore, at least two links must connect nodes between which commodities can be transported in either direction. The nodes are enumerated in any order, consecutively beginning with one, as illustrated by the graph in Fig. 1.

Let p_d^i and p_s^i be n vectors denoting the unit demand price and the unit supply price of n commodities at node i, and let $p^{ij} \in R^n$ denote the cost of transporting each of these n commodities over link ij. Affine relations are assumed between prices and quantities; i.e.,

$$p^i = A^i x^i + a^i \quad \text{for each node } i \in N, \tag{5}$$

where $p^i = (p_d^i, p_s^i)^T$, A^i, and a^i are given constant matrices and vectors (which arose in [4] from inverting the difference between given supply and demand quantities, originally expressed as affine functions of price), and x^i is an n-dimensional vector representing the excess quantity of each commodity produced by node i. It is also assumed that

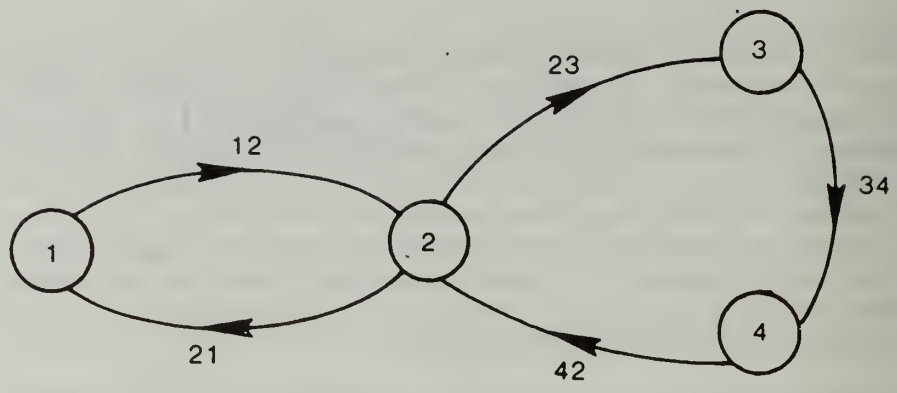

FIGURE 1. Sample transshipment network.

(6) $$p^{ij} = A^{ij}x^{ij} + a^{ij} \text{ for each link } ij \in L,$$

where A^{ij} and a^{ij} are given constant matrices and vectors (which arose from describing the transport prices as functions of transport volumes), and $x^{ij} \in R^n$ denotes the quantity of n commodities transported over link ij.

The quantities are constrained by the nonnegativity condition

(7) $$x^{ij} \geqq 0 \text{ for each link } ij \in L$$

and the commodity conservation condition

(8) $$x^i = \sum_{j \in N} x^{ij} - \sum_{j \in N} x^{ji} \text{ for each node } i \in N,$$

where $x^{ij} = 0$ if $ij \notin L$. Note that although x^{ij} is nonnegative by virtue of the choice of link direction, the components of x^i might be positive or negative depending on whether node i is a net exporter or net importer of a particular commodity.

The price stability condition on p leads to the following relationship:

(9) $$p^{ij} + p_s^i \geqq p_d^j \text{ for each link } ij \in L.$$

To see this, assume that Eq. (9) is violated for some commodity c. As a consequence some economic agent would find it profitable to purchase as much of commodity c as possible at node i and transport it over link ij for resale at node j. This would clearly be an economically unstable situation.

The final relationship needed to establish equilibrium is the complementarity condition

(10) $$< x^{ij},\ (p_d^j - p^{ij} - p_s^i) > = 0 \text{ for each link } ij \in L.$$

This condition is imposed to ensure that no positive flow x^{ij} will occur on a link if the cost $p^{ij} + p_s^i$ of a commodity at node j exceeds the price p_d^j which a consumer is willing to pay.

To conform with the notation developed in Section 1.0, a function $g^{ij} \in R^n$ equivalent to the excess demand function but now expressed in terms of prices rather than quantities will be defined by the following expression:

$$g^{ij} \doteq p_d^j - p^{ij} - p_s^i \text{ for each link } ij \in L.$$

Through the appropriate substitutions the solution (x,p) to the economic equilibrium conditions (5-10) can be described entirely in terms of the solution (x,g) to the linear complementarity conditions

(11) $$x \geqq 0\ ,\ \langle x,g\rangle = 0,\ g \leqq 0$$

(12) $$g = -Mx - \mathrm{v}$$

where x, g, and v are vectors equal in size to the number of links times the number of commodities, and M is a square matrix of comparable dimension whose components are given in Fig. 2. The constant v follows from the substitution of Eqs. (5) and (6) into Eq. (10) and is given by

$$\mathrm{v}^{ij} = a_d^j - a^{ij} - a_s^i \text{ for each link } ij \in L.$$

$[M \| \mathrm{v}] =$

$A_1+A_2+A^1$	$-(A_1+A_2)$	$-A_2$	0	A_2	v^{12}
$-(A_1+A_2)$	$A_1+A_2+A^2$	A_2	0	$-A_2$	v^{21}
$-A_2$	A_2	$A_2+A_3+A^3$	$-A_3$	$-A_2$	v^{23}
0	0	$-A_3$	$A_3+A_4+A^4$	$-A_4$	v^{34}
A_2	$-A_2$	$-A_2$	$-A_4$	$A_2+A_4+A^5$	v^{42}

FIGURE 2. Constituents of matrix for sample network.

For purposes of illustration, the 2-commodity, 5-link network shown in Fig. 1 has been considered for analysis. The reworked data for this problem are displayed in Fig. 3. If conditions (11) and (12) are now put into the format of problem (S), we get a problem of the form

(A)
$$\min_{x,w} \sum_{i=1}^{10} (\min(0,w_i) + x_i)$$
$$w_i - M_i x + x_i = \mathrm{v}_i$$
$$-M_i x \leqq \mathrm{v}_i \qquad i = 1, 2, \ldots, 10$$
$$x \geqq 0$$

where M_i is the ith row of M.

The algorithm that is used for the computations does not solve the original problem (S), but constructs an approximate problem to solve by replacing each of the associated functions with their piecewise linear convex envelopes. A related problem is simultaneously introduced which gives a sharper underestimate of the optimal value of the approximating problem than does the convex envelope problem. It is this related problem that the branch and bound procedure solves first to get estimates on the optimal value of the approximating problem, and to set up new problems if the estimates do not yield a global solution.

$[M \| v] =$

4	-1	-2	1	-1	1	0	0	1	-1	-1
2	3	-2	-2	0	-1	0	0	0	1	2
-2	1	3	0	1	-1	0	0	-1	1	-1
-2	-2	2	3	0	1	0	0	0	-1	-5
-1	1	1	-1	3	-3	-1	1	-1	1	-2
0	-1	0	1	2	3	-1	-1	0	-1	1
0	0	0	0	-1	1	2	-2	0	2	0
0	0	0	0	-1	-1	4	2	-2	0	-3
1	-1	-1	1	-1	1	0	2	2	-3	2
0	1	0	-1	0	-1	-2	0	4	2	4

FIGURE 3. Data for affine network.

The functions defining the constraint region of problem (A) are all linear and hence convex, and therefore will not be replaced in the approximate problem. The functions associated with the nonlinear variables w_i in the objective function (i.e., $\min(0, w_i)$, $i = 1, \ldots, n$) are piecewise linear, but concave, and will be replaced in the approximate problem by their convex envelopes, which in this case are straight lines. This is illustrated in Fig. 4.

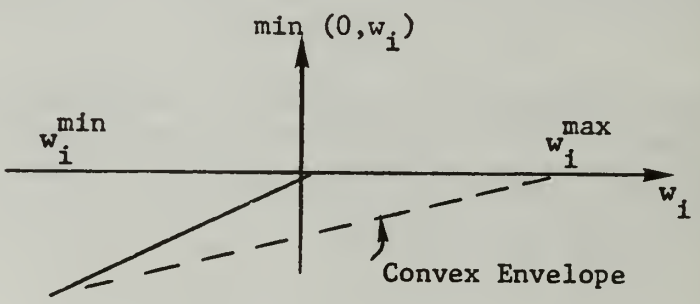

FIGURE 4. Convex envelope of $\min(0, w_i)$.

The branch and bound technique proceeds to divide the domain of these functions into pieces corresponding to their linear segments and separately solves the set of related problems in which the nonlinear variables are respectively limited. When every function is piecewise linear, as is the case with problem (A), we get an exact solution to the original problem.

It is customary with branch and bound methods to describe the algorithm in terms of a branch and bound tree. The nodes of the tree correspond to the related linear subproblems, while the branches of the tree correspond to the set on which the branching variables are defined. A solution is obtained when the best upper bound at any node is equal to zero. That is, any feasible point yielding a subproblem value of zero necessarily satisfies the equilibrium

conditions. For problem (A), no tree developed because the solution was obtained on the first iteration of the algorithm. The optimal vector x is given by

$$x^{12} = (0.2353,\ 0.7059)\ x^{21} = (0.0,\ 2.2941)\ x^{23} = (1.5294,\ 0.0)$$
$$x^{34} = (1.0098,\ 0.2151) \qquad x^{42} = (0.0,\ 0.0).$$

The formulation of problem (A) required the addition of 10 auxiliary variables to the original set of 10 linear variables. The former were each divided into two intervals for the purposes of branching. This division, corresponding to the segments of the piecewise linear functions defined for these variables, implied that any branch and bound tree produced by the algorithm could be at most 10 branches deep and that no variable could appear more than once along any path. In theory, it might have been necessary to solve up to $2^{11} - 1$ subproblems before reaching a solution; however, the fact that the first subproblem produced an equilibrium point underscores the computational efficiencies that result from having available at the outset a means of independently checking each iteration for convergence.

Each subproblem solved by MOGG is a linear program. When the excess demand functions are affine, the solution vector necessarily yields a feasible point to the original problem. The upper bound associated with this feasible point will always be greater than or equal to zero, but generally not correspond to an equilibrium solution. Mangasarian [18] has shown that the linear complementarity problem is equivalent to a linear program whose cost coefficients are dependent upon the structure of M. The similarity between these linear programs and those set up by MOGG when M meets certain conditions admits the possibility that MOGG will produce an equilibrium point on the first iteration. Although these conditions might logically arise in some economic contexts, they were not present in this example and, hence, did not influence the rate of convergence.

3.2 A Piecewise Linear Market

This example [7] provides a simple explanation of how a competitive market operates. As is common in microeconomic theory, we will distinguish among individuals according to the economic functions that they perform or on the basis of the kinds of decisions they make. Thus, a consumer is an individual (or unit) that consumes commodities and supplies inputs to production. The role of the consumer may be defined as that of choosing from among the alternative commodity bundles available to him. Similarly, a producer is an individual (or group) that utilizes inputs to produce commodities. The role of the producer may be characterized as that of choosing from among the alternative input-output patterns available to him. The same individual might appear in the economy both as a consumer and as a producer. Once the choices are made, a state of the economy is defined.

Under certain assumptions (see Quirk and Saposnik [22]), for a market that contains n commodities, m consumers, and l producers, the aggregated (net) amounts of commodities demanded and supplied for any vector of prices can be determined by a simple summation of the amounts demanded and supplied by individual consumers and producers. Thus, given the price vector p, where $p = (p_1, p_2, \ldots, p_n)$, we can write $x_{ij}(p)$ as the amount of the i th commodity consumed (or supplied as an input in production) by the j th individual at the set of prices given by p ; and $y_{ik}(p)$ as the amount of the i th commodity produced (or used up as an input in production) by the k th firm at the set of prices given by p. Then, the aggregate (net) consumption of commodity i by consumers is given by

$$x_i(p) = \sum_{j=1}^{m} x_{ij}(p), \quad i = 1, 2, \ldots, n$$

and the aggregate (net) production of commodity i by producers is given by

$$y_i(p) = \sum_{k=1}^{l} y_{ik}(p), \quad i = 1, 2, \ldots, n.$$

We then define $x(p)$ and $y(p)$ as point-to-point mappings from R^n into itself. In the absence of any initial endowment the excess demand function can be written as $g(p) = x(p) - y(p)$.

The sample economy under consideration contains three commodities, three consumers, and three producers. The associated supply and demand functions are assumed to be piecewise linear, and are given in graphic form in Figs. 5 and 6. To conform with the presentation in [7], the equilibrium quantity rather than the equilibrium price will be computed. The following notation will be used:

$p_{s_{ij}}$ = j th producer's supply price for commodity i,

$p_{d_{ij}}$ = j th consumer's demand price for commodity i,

x_i = quantity of i th commodity consumed,

y_i = quantity of i th commodity produced,

where $i, j = 1, 2, 3$, $p_{s_i} = \sum_j p_{s_{ij}}$ is a function of the consumption variable x and $p_{d_i} = \sum_j p_{d_{ij}}$ is a function of the production variable y.

An equilibrium point will exist if the following conditions are satisfied:

(13) $x \geqq 0, \ y \geqq 0,$

(14) $x - y = 0,$

(15) $p_d - p_s \leqq 0,$

(16) $<x, \ p_d - p_s> = 0,$

where p_s and p_d are the three-dimensional market supply price and market demand price vectors. The first condition assures feasibility; the second condition clears the market; the third condition assures price consistency by requiring the excess demand function to be less than or equal to zero; and the fourth condition is Walras' Law and reflects the following circumstances: if x_i, the quantity of the i th commodity being purchased, was positive and if the producers' supply price p_{s_i} was greater than the consumers' demand price p_{d_i}, then the producers would be losing money and would begin to lower $y_i (= x_i)$ to zero. Such a situation would be economically unstable.

Conditions (13)—(16) can now be written as a minimization problem in the form of problem (P):

$$\min\{<-x, \ (p_d - p_s)>\}$$

$$p_d - p_s \leqq 0$$

$$x \geqq 0$$

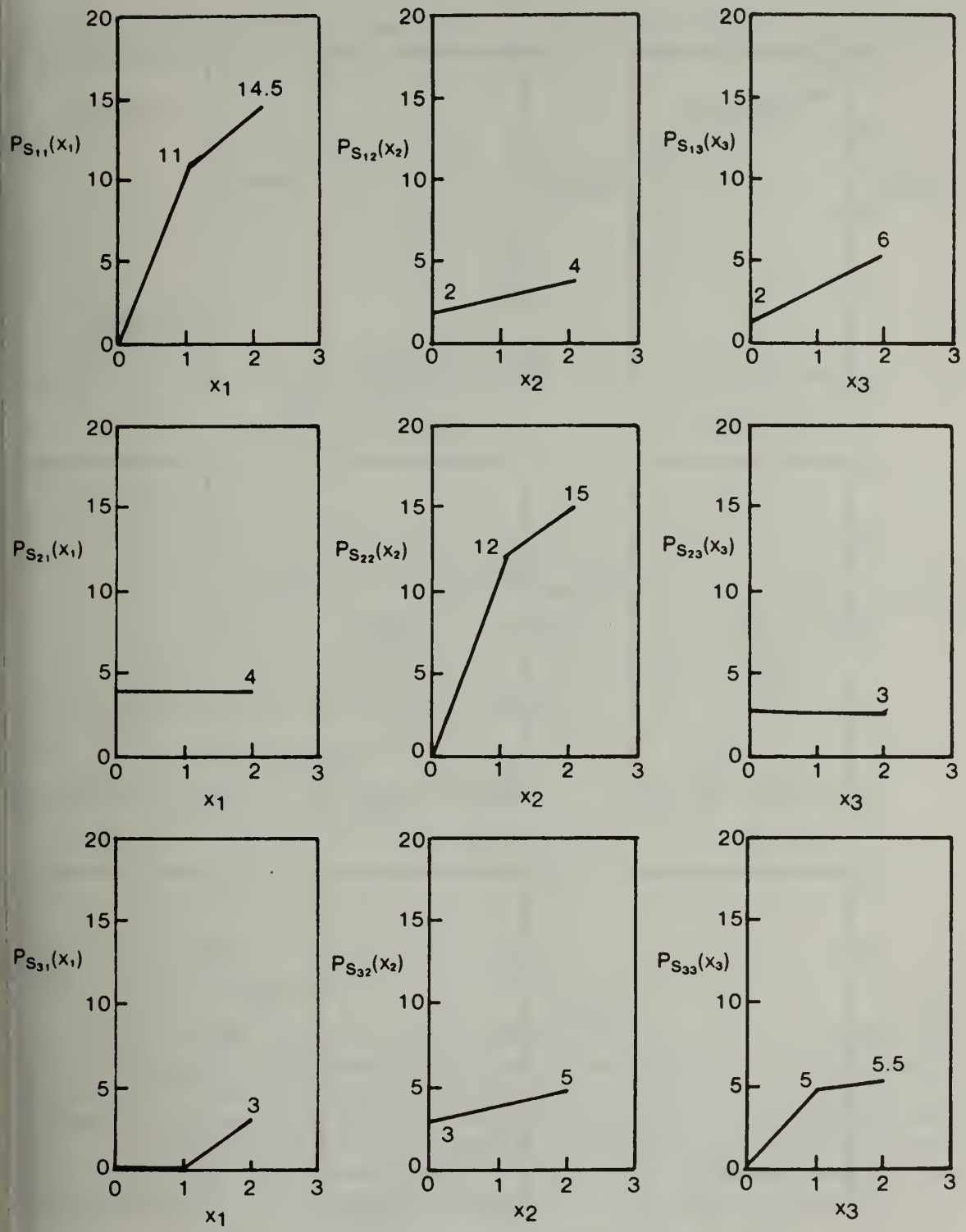

FIGURE 5. Supply functions for piecewise linear market.

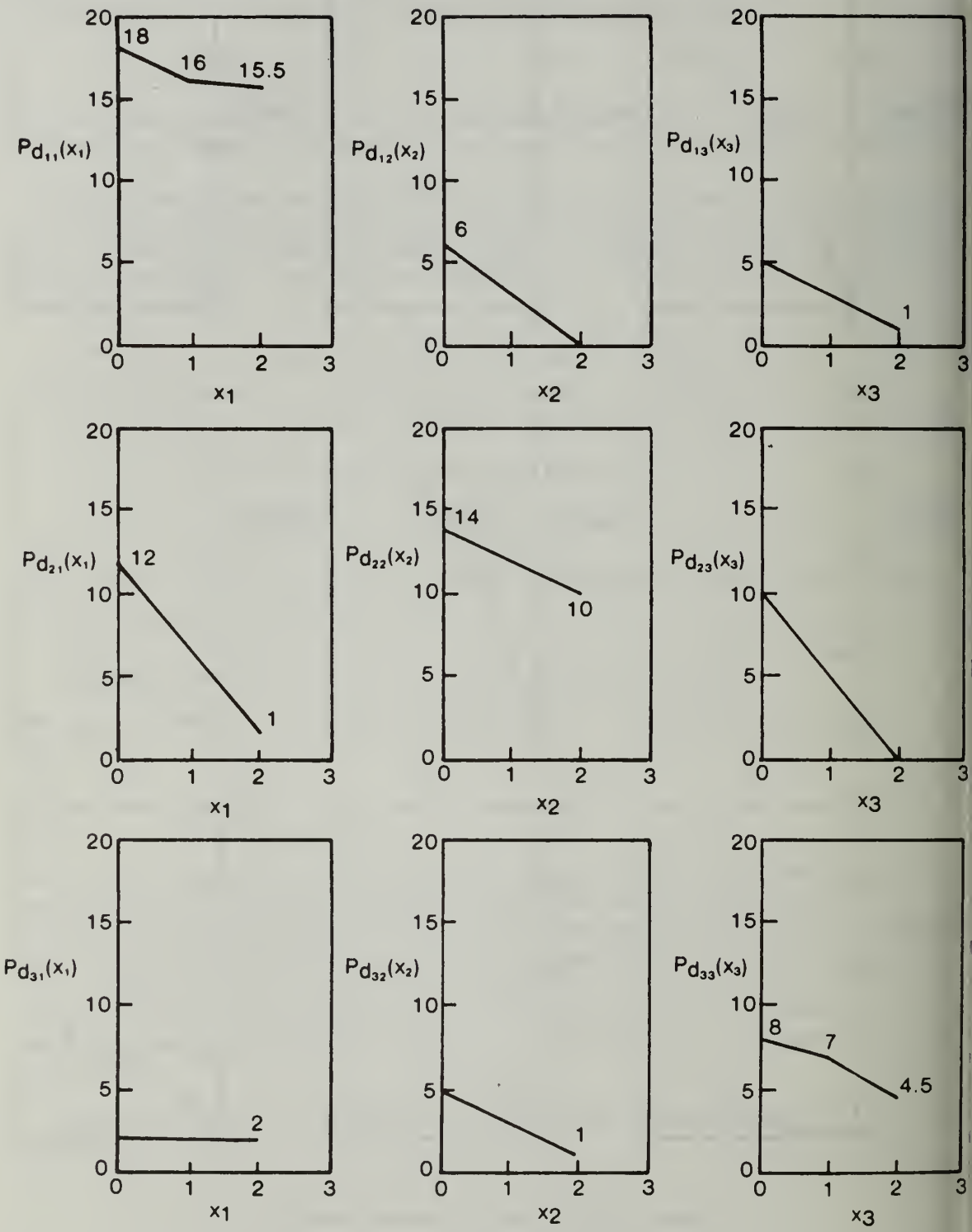

FIGURE 6. Demand functions for piecewise linear market.

In order to recast this problem in the form of problem (S), the consumption and production data given in Figs. 5 and 6 must be aggregated over their respective agents to obtain the market demand and supply curves p_d and p_s. This has been done for each of the three commodities.

COMMODITY 1:

$$p_{d_1} = \begin{cases} -2x_1 + 18; & x_1 \leqq 1 \\ -0.5x_1 + 16.5; & x_1 > 1 \end{cases} + (6 - 3x_2) + (5 - 2x_3)$$

$$p_{s_1} = \begin{cases} 11x_1; & x_1 \leqq 1 \\ 7.5x_1 - 3.5; & x_1 > 1 \end{cases} + (2 + x_2) + (2 + 2x_3)$$

COMMODITY 2:

$$p_{d_2} = (12 - 5x_1) + (14 - 2x_2) + (10 - 5x_3)$$

$$p_{s_2} = 4 + \begin{cases} 12x_2 & ; x_2 \leqq 1 \\ 3x_2 + 9; & x_2 > 1 \end{cases} + 3$$

COMMODITY 3:

$$p_{d_3} = 2 + (5 - 2x_2) + \begin{cases} -x_3 + 8; & x_3 \leqq 1 \\ -2.5x_3 + 9.5; & x_3 > 1 \end{cases}$$

$$p_{s_3} = \begin{cases} 0; & x_1 \leqq 1 \\ 3x_1 - 3; & x_1 > 1 \end{cases} + (3 + x_2) + \begin{cases} 5x_3; & x_3 \leqq 1 \\ 0.5x_3 + 4.5; & x_3 > 1 \end{cases}$$

The minimization problem in its separable form becomes

(B)
$$\min_{x,w} \sum_{i=1}^{3} \{\min(0, w_i) + x_i\}$$

$$w_i + p_{d_i} - p_{s_i} + x_i = 0 \qquad i = 1, 2, 3$$

$$p_{d_i} - p_{s_i} \leqq 0$$

$$x \geqq 0$$

where p_{d_i} and p_{s_i} are defined above.

Each of the six variables in problem (B) is nonlinear, the first three (w_1, w_2, w_3) corresponding to the auxiliary variables and the second three (x_1, x_2, x_3) to the original problem variables. The associated functions are piecewise linear and contain at most one break point. This means that the branch and bound tree can be at most six nodes deep and that a maximum of $2^7 - 1$ subproblems might have to be solved. Once again though, the algorithm converges on the first iteration. The computed best upper bound for the first subproblem is zero and hence the solution.

If the algorithm is permitted to run past this point until its usual termination conditions are met a total of 19 subproblems will be solved. Figure 7 depicts the resultant branch and bound tree which serves to illustrate both the advantage of knowing the optimal value at the outset and the amount of work required to search for alternative solutions. The known results

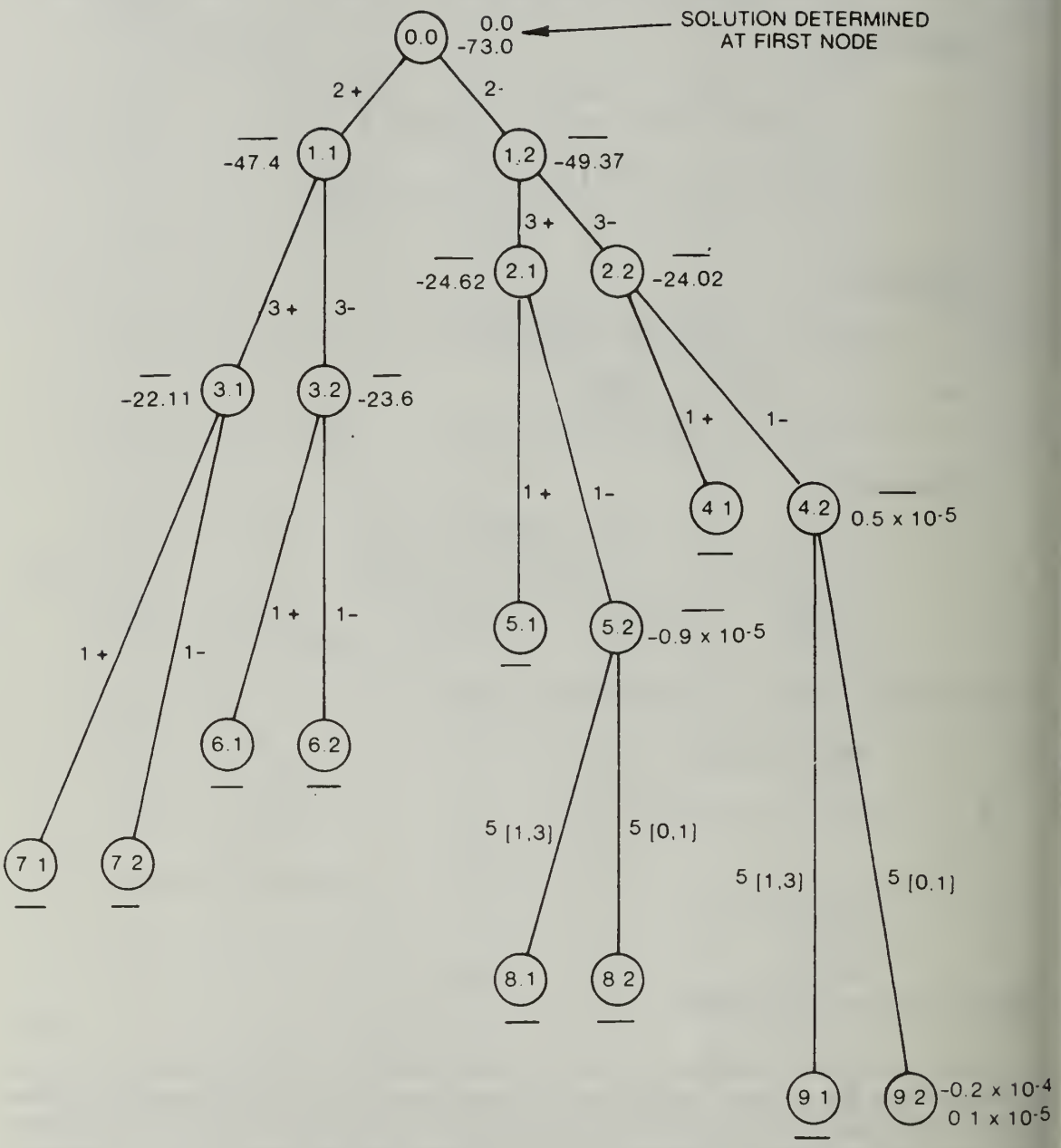

FIGURE 7. Branch and bound tree for piecewise linear market when forced past solution.

are corroborated at node 9.2 where roundoff errors have produced a best upper bound of -0.2×10^{-4} and a best lower bound of 0.1×10^{-5}, a minor contradiction. The two numbers adjacent to each node represent the best upper and lower bounds for that subproblem. A bar in the place of the best upper bound indicates that no corresponding feasible point to the approximate problem exits. The numbers along the branches refer to the branching variables associated with the preceding node, and the + and − signs indicate whether the particular auxiliary variable was permitted to range over the set of positive real numbers or negative real numbers, respectively. The bars appearing below the nodes indicate that either the lower bounds of the associated subproblems are all greater than the current best upper bound or that they are infeasible and, therefore, cannot contain the solution.

In terms of the actual variables, the solution vectors are $x^* = (2, 1, 1)$ and $w^* = (-2, -1, -1)$. From the equality constraints in problem (B), it can be seen that $x^* + w^* = 0$ whenever the corresponding excess demand functions are binding. By tracing the convergent path backwards from node 9.2 to node 0.0 we see that the branches that fall along this path correspond to the nonpositive orthant of w. The subscripts attached to the branching variables in the tree denote the (closed) intervals over which the original problem variables are defined for all subsequent subproblems.

3.3 A General Market

The separable programming algorithm works by first replacing each of the original problem functions with their piecewise linear convex envelopes, and then creating a new problem to solve as an ultimate approximation. From this approximate problem a series of convex subproblems issue that are set up and solved under the branch and bound philosophy. If the original functions are all piecewise linear (but not necessarily convex), then solving the aggregate of subproblems is tantamount to solving the original problem exactly. Such was the case in both the first and second examples. In this example, four of the piecewise linear functions in Fig. 5 and 6 have been replaced with smooth counterparts. The new functions were constructed to pass through the points $(1,\cdot)$ and $(2,\cdot)$, and are given by

$$p_{d_{11}} = 16.516\, e^{(-0.03175 x_1)},$$

$$p_{d_{33}} = -0.75\, x_3^2 - 0.25\, x_3 + 8,$$

$$p_{s_{11}} = -3.75\, x_1^2 + 14.75\, x_1,$$

$$p_{s_{22}} = 17.31234 \log(x_2^{0.463} + 1).$$

Substituting these functions for the originals in problem (B) leads to a new minimization problem that can be written as

$$\text{(C)} \qquad \min_{w,x} \sum_{i=1}^{3} (\min(0, w_i) + x_i)$$

subject to

$$w_1 + 3.75\, x_1^2 - 13.75 x_1 + 16.516\, e^{(-0.03175 x_1)} - 4x_2 - 4x_3 = -7$$

$$w_2 - 5x_1 - 17.31234 \log(x_2^{0.463} + 1) - x_2 - 5x_3 = -29$$

$$w_3 - \begin{cases} 0; & x_1 \leqq 1 \\ 3x_1 - 3; & x_1 > 1 \end{cases} - 3x_2 - \begin{cases} 0.75x_3^2 + 4.25x_3 - 8; & x_3 \leqq 1 \\ 0.75x_3^2 - 0.25x_3 - 3.5; & x_3 > 1 \end{cases} = -4$$

$$3.75\, x_1^2 - 14.75\, x_1 + 16.516\, e^{(-0.03175\, x_1)} - 4x_2 - 4x_3 \leqq -7$$

$$-5x_1 - 17.31234\ \log(x_2^{0.463} + 1) - 2x_2 - 5x_3 \leqq -29$$

$$\begin{Bmatrix} 0; & x \leqq 1 \\ 3x_1 - 3; & x_1 > 1 \end{Bmatrix} - 3x_2 - \begin{Bmatrix} 0.75x_3^2 + 5.25x_3 - 8; & x_3 \leqq 1 \\ 0.75x_3^2 + 0.75x_3 - 3.5; & x_3 \geqq 1 \end{Bmatrix} \leqq -4$$

$$x_1 \geqq 0,\ x_2 \geqq 0,\ x_3 \geqq 0.$$

The number of cuts required to approximate a piecewise linear function exactly is equal to the number of segments constituting that function. When the function is smooth, it cannot be represented exactly by a finite number of linear segments but can be approximated with arbitrary precision by increasing the number of cuts. In this example, six cuts were made in the original problem variables (x_1,x_2,x_3) over the closed interval [0,3]. Because the cuts were evenly spaced every half integer, and the graphs of the smooth functions pass through the solution points of problem (B), it is reasonable to expect that the solution to problem (C) would coincide with one or more of these points. This indeed was the case: the identical solution $x_1^* = (2,1,1)$ resulted for problem (C). The associated branch and bound tree is shown in Fig. 8. The algorithm is seen to have converged in the tenth stage at node 10.1 after 22 subproblems have been solved. This contrasts with the first two examples where the solution occurred on the first iteration; however, in each of these three problems, the first feasible point produced by the algorithm resulted in the solution. Finally, we observe from Fig. 8 that at the tenth stage, the best upper bound and the best lower bound are nominally equal, implying that the general conditions for optimality have been satisfied, so the algorithm is terminated. If an equilibrium point had not yet been found at this stage, it would have been reasonable to conclude that none existed for the given model. The other two equilibrium points were not uncovered.

3.4 Network Traffic Flow

The model of the road system considered here derives from the notion that there exists a large community of users, each of whom takes the quickest route available, given the actions of other users. The number of trips taken is assumed to depend on the time required to make a trip, while the travel time on a particular road is assumed to depend on traffic volume. The example that we will investigate was studied by Asmuth [3], who used stationary point theory in conjunction with the Eaves-Saigal algorithm [9] to obtain a solution. As will be seen, the traffic flow problem closely resembles the multicommodity network presented in the first example.

To formulate the model, consider a directed network (N,A) with nodes i in N and arcs ij in A. For each arc ij, we are given a delay function f_{ij} which expresses travel time on arc ij as a function of the traffic flows on the arcs of the network. The travel time along arc ij will necessarily depend on the flow on that arc, but may as well depend on flows along other arcs. For example, a two-way street could be modeled as a pair of opposing arcs where the flow on one of the arcs might affect the travel time on the other.

For each pair of distinct nodes i and k we are also given a travel demand function $g_{i,k}$ which expresses demand for travel from i to k as a function of travel times between nodes on the network. Demand for travel from i to k will depend on travel time from i to k as well as on travel times between other pairs of nodes; for instance, for i to some alternate destination.

Numerous solution procedures have been proposed for computing equilibrium traffic flows and travel times on the network. When f is integrable and g has an inverse which is integrable, the usual approach has been to reformulate the equilibrium problem as a convex programming problem. These conditions will be met if each f_{ij} depends only on the total flow along arc ij,

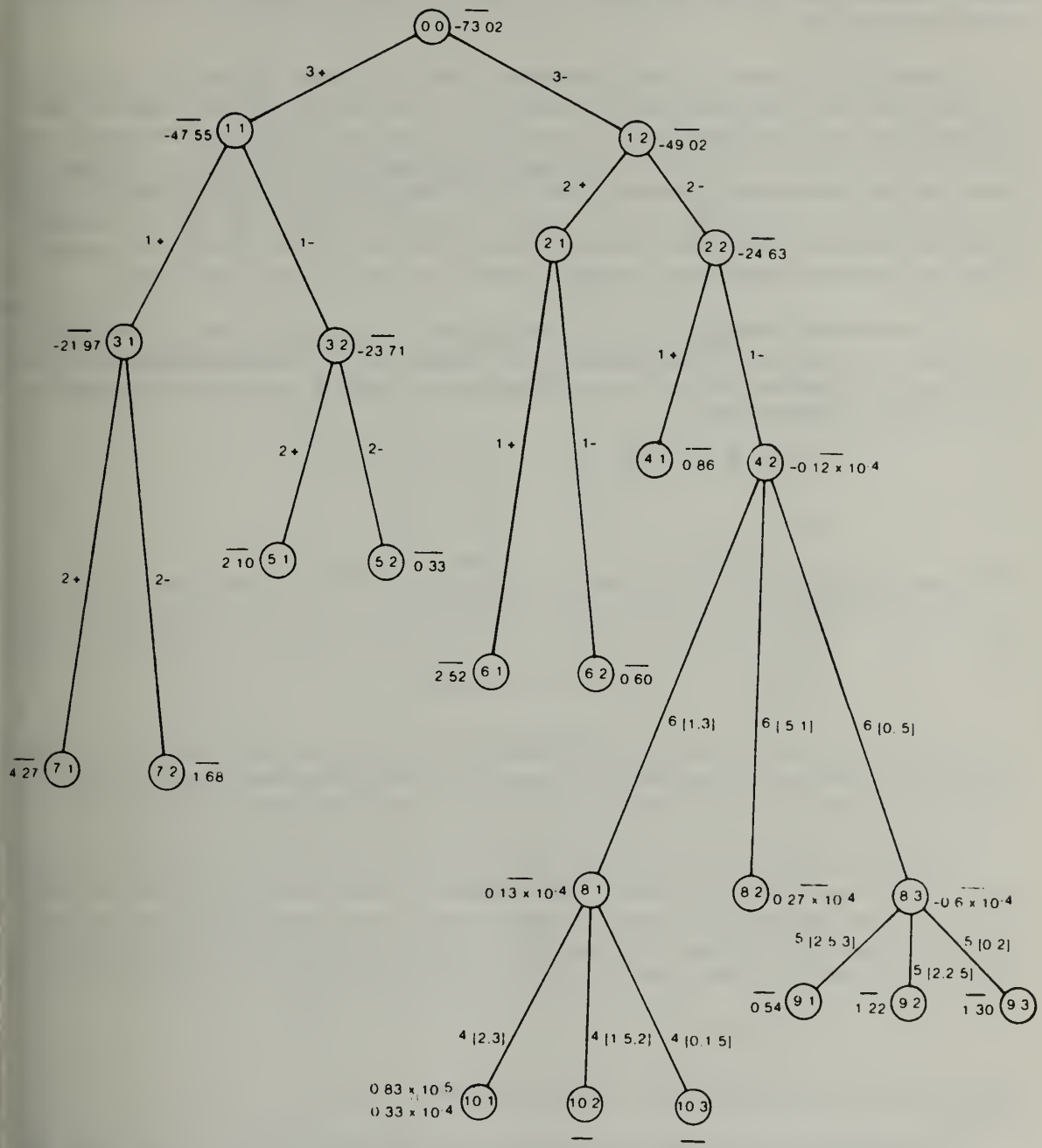

FIGURE 8. Branch and bound tree for general market.

and each $g_{i,k}$ is monotonically decreasing and depends only on the travel time from node i to node k. Beckman [6], Florian and Nguyen [14], and others have addressed this problem under comparable conditions.

In practice, the demand and delay functions f and g are at best empirical fits and can be endowed with these or any other restrictions which may seem useful. Asmuth's approach does not depend on such restrictions, but only requires that the delay functions f_{ij} be positive on each arc, that the travel demand functions $g_{i,k}$ be nonnegative and bounded for each pair of nodes, and that the network be complete; that is, a directed path must exist from every node to every other node. From a strictly analytic point of view, we will require only that it be possible to put the functions f and g into a separable form. However, if the model is to accurately reflect the properties of the system it might be desirable to adopt the above restrictions.

The mathematical conditions for a user equilibrium are presented below. The travel time from node i to node k will be written as $t_{i,k}$ and the flow on arc ij with destination k will be written as $y_{ij,k}$. It will be said that the travel time vector t and flow vector y are in equilibrium if the following conditions hold:

$$g_{i,k}(t) = \sum_j y_{ij,k} - \sum_j y_{ji,k} \qquad i \neq k,\ i,\ k \in N \tag{17}$$

$$y \geqq 0 \tag{18}$$

$$t_{i,k} \leqq f_{ij}(y) + t_{j,k} \qquad i \neq k,\ ij \in A,\ k \in N \tag{19}$$

$$t_{k,k} = 0 \qquad k \in N$$

$$y_{ij,k}(f_{ij}(y) + t_{j,k} - t_{i,k}) = 0 \qquad i \neq k,\ ij \in A,\ k \in N \tag{20}$$

$$y_{ij} = \sum_k y_{ij,k} \qquad ij \in A \tag{21}$$

Condition (17) is the conservation-of-flow equation. It says that the traffic leaving node i with destination k is the sum of the traffic arriving at node i with destination k and the traffic originating at i with destination k. Condition (18) says that traffic flows cannot be negative.

Conditions (19) and (20) require that traffic flow be by the fastest route available. In condition (19) we require that $t_{i,k}$ not exceed the minimum travel time from i to k, given the flows y on the network; condition (20) limits the traffic to those routes which achieve this minimum travel time. Together, conditions (19) and (20) imply the principle of minimum travel time. This says that if any traffic flows from i to k, that is, if $\sum_j y_{ij,k} > 0$, then

$$t_{i,k} = \min_j (f_{ij}(y) + t_{j,k}).$$

Equation (21) relates the basic flows to the total arc flows.

It may be useful to think of this system as a multicommodity network, where all of the traffic destined for a particular node k is a separate commodity, all of which must be shipped to node k via the network. In this way $g_{i,k}(t)$ is the amount of commodity k which must travel from node i to node k. This trip will traverse a path of arcs from i to k.

Conditions (17-21) can be put in the form of problem (P) as follows:

$$\min_{\substack{y \geqq 0 \\ t \geqq 0}} \sum_{k \in N} \sum_{\substack{ij \in A \\ i \neq k}} < y_{ij,k},\ (-f_{ij}(y) - t_{j,k} + t_{i,k}) >$$

subject to

$$-f_{ij} - t_{j,k} + t_{i,k} \leqq 0 \qquad i \neq k,\ ij \in A,\ k \in N,$$

$$g_{i,k} - \sum_{j} y_{ij,k} + \sum_{j} y_{ji,k} = 0 \qquad i \neq k,\ i,\ k \in N,$$

$$y_{ij} - \sum_{k} y_{ij,k} = 0 \qquad ij \in A.$$

Rewriting this problem in the form of problem (S) we get

$$\min_{\substack{v \geqq 0 \\ t \geqq 0 \\ w}} \sum_{k \in N} \sum_{\substack{ij \in A \\ i \neq k}} \{\min(0, w_{ij,k}) + y_{ij,k}\}$$

subject to

$$w_{ij,k} - f_{ij} - t_{j,k} + t_{i,k} + y_{ij,k} = 0 \qquad i \neq k,\ ij \in A,\ k \in N,$$

$$-f_{ij} - t_{j,k} + t_{i,k} \leqq 0 \qquad i \neq k,\ ij \in A,\ k \in N,$$

$$g_{i,k} - \sum_{j} y_{ij,k} + \sum_{j} y_{ji,k} = 0 \qquad i \neq k,\ i,\ k \in N,$$

$$y_{ij} - \sum_{k} y_{ij,k} = 0 \qquad ij \in A.$$

The following sample problem is from [4] and is based on the directed network shown in Fig. 9. Here $N = \{1, 2, 3\}$ and $A = \{12, 21, 23, 31\}$, where arcs 12 and 21 represent a two-way street. The delay functions are

$$f_{12}(y) = 10 + e^{(v_{12} - 10)} + 1.25 \log(y_{21} + 1.0),$$

$$f_{21}(y) = 10 + e^{(v_{21} - 10)} + 1.25 \log(y_{12} + 1.0),$$

$$f_{23}(y) = 4 + e^{(v_{23} - 12)},$$

$$f_{31}(y) = 4 + e^{(v_{31} - 20)},$$

where

$$y_{ij} = \sum_{k} y_{ij,k},$$

and the travel demand functions are

$$g_{1,2}(t) = \frac{80}{t_{1,2} + 1},$$

$$g_{1,3}(t) = \frac{120}{t_{1,3} + 1},$$

$$g_{2,1}(t) = \begin{cases} \dfrac{40}{t_{2,1} + 1} & \text{if } t_{2,1} \geqq t_{2,3}, \\ \dfrac{100}{t_{2,1} + 1} & \text{if } t_{2,1} \leqq t_{2,3}, \end{cases}$$

$$g_{2,3}(t) = \begin{cases} \dfrac{80}{t_{2,3} + 1} & \text{if } t_{2,1} \geqq t_{2,3}, \\ \dfrac{20}{t_{2,3} + 1} & \text{if } t_{2,1} \leqq t_{2,3}, \end{cases}$$

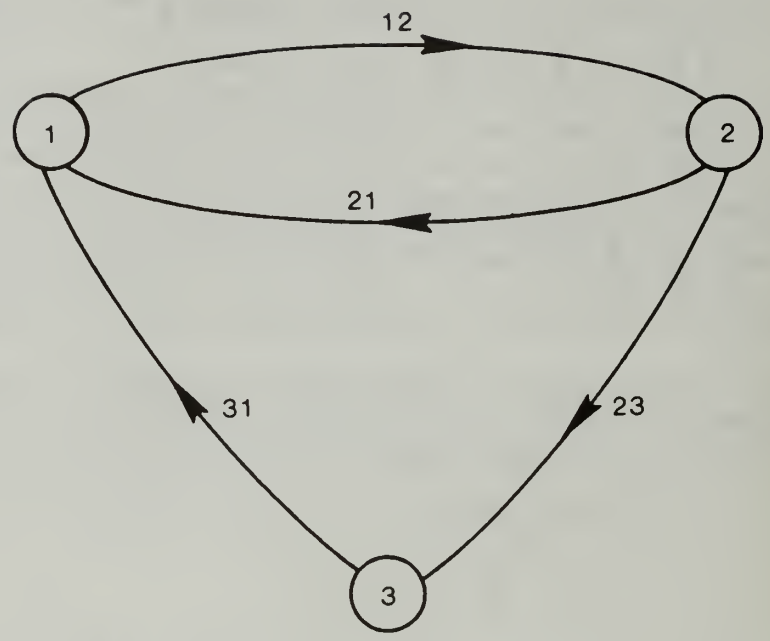

FIGURE 9. Sample traffic network.

$$g_{3,1}(t) = \frac{60}{t_{3,1} + 1},$$

$$g_{3,2}(t) = \frac{100}{t_{3,2} + 1}.$$

When more than one function value is given at a particular point (e.g., $t_{2,1} = t_{2,3}$), the value of g is the average of the two values. In this case some of the travelers from node 2 will go to either 1 or 3 depending on which is closest. If the travel times are equal then those travelers who want to go to either 1 or 3 will be divided between the two destinations.

In their present form, the demand functions $g_{2,1}$ and $g_{2,3}$ exhibit an implicit dependency on the travel times $t_{2,1}$ and $t_{2,3}$ and therefore must be made separable before the equilibrium problem can be solved. Although this cannot be done explicitly, the desired result can be achieved by considering the following three disjoint partitions of t:

$$t_{2,1} < t_{2,3}, \quad t_{2,1} = t_{2,3}, \quad t_{2,1} > t_{2,3}.$$

The mathematical program associated with each of these partitions comprises 26 variables and 27 constraints. Of the 26 variables, 12 are of the type required to achieve separability of the functions while the remainder are defined in the original problem statement.

The solution was uncovered in the third partition at the 84th stage after 168 subproblems had been solved, and once again, coincided with the first feasible point found. The resultant branch and bound tree is not displayed because of its extensive length, but the final computations are highlighted in Table 1 along with the results obtained by the Eaves-Saigal algorithm. The minor discrepancies observed between the variable and functional values computed by either method can be attributed to the grid size superimposed on the algorithm and are, hence,

TABLE 1 — *Results for Traffic Network Problem*

	Eaves-Saigal Algorithm		MOGG	
i,k	$t_{i,k}$	$g_{i,k}(t)$	$t_{i,k}$	$g_{i,k}(t)$
1,2	19.30	3.94	19.45	3.91
1,3	28.43	4.08	28.58	4.06
2,1	13.22	2.81	13.21	2.81
2,3	9.13	7.90	9.13	7.90
3,1	4.09	11.79	4.09	11.79
3,2	23.38	4.10	23.54	4.07
ij,k	$y_{ij,k}$	f_{ij}	$y_{ij,k}$	f_{ij}
12,2	8.04	19.30	7.99	18.68
12,3	4.08	19.30	4.06	18.68
21,1	1.15	13.22	1.21	13.21
21,3	0.00	13.22	0.00	13.21
23,1	1.66	9.13	1.60	8.76
23,3	11.97	9.13	11.96	8.76
31,1	13.46	4.09	13.41	4.08
31,2	4.10	4.09	4.08	4.08

subject to control. Finer resolution is strictly a matter of increasing the number of grid points prescribed for the original nonlinear variables and solving a proportionately larger problem.

4.0 CONCLUSIONS

The computation of equilibria plays a major role in the analysis of economic and transport systems. Whenever the equilibrium problem can be formulated as a set of complementarity equations, we have shown for those cases where the original functions are implicitly separable, that nonconvex programming can be used to obtain a solution to either problem. A general algorithm based on branch and bound techniques was adapted to perform the equilibrium computation. Unlike the usual nonconvex program though, where the solution is recognized only when equality is achieved between the best upper and best lower bounds, an independent check can be made for the solution at each iteration because the best upper bound is known at the outset. As our computational experience demonstrates, this enhancement markedly increases the efficiency of the algorithm.

However, the fact that a numerical procedure will terminate with the correct answer in a finite number of iterations is no guarantee that it will be of any practical use. The combination of method and algorithm under study derives its tentative usefulness from the observation that for most problems investigated, convergence occurred in a far smaller number of iterations than theoretically possible. The results have been especially encouraging for problems of larger dimensions; and in all cases, the equilibrium solution coincided with the first feasible point found by MOGG.

The affine equilibrium problem or linear complementarity problem holds a particular interest because of its unique structure and implicit relationship to an equivalent linear program. Because of the similarity between the first subproblem set up by MOGG and this linear program, immediate solutions are often obtainable from MOGG at little extra cost. In fact, the

additional work required to determine the equivalent linear program, even for relatively small problems, is often more burdensome and more computationally expensive than permitting MOGG to run beyond its first subproblem to a point of convergence. A further and decided advantage of MOGG is that it will solve all affine equilibrium problems regardless of their matrix structure. By contrast, the majority of alternative procedures available are limited in their application to a number of special cases which do not necessarily arise in practice.

REFERENCES

[1] Arrow, K.J. and G. Debreu, "Existence of an Equilibrium for a Competitive Economy," Econometrica *22*, 265-90 (1954).

[2] Arrow, K. and F. Hahn, *General Competitive Analysis*, Holden-Day Publishing Company, San Francisco, Calif. (1971).

[3] Asmuth, R.L., "Traffic Network Equilibria," Technical Report SOL 78-2. Department of Operations Research, Stanford University (1978).

[4] Asmuth, R.L., B.C. Eaves and E.L. Peterson, "Studying Economic Equilibria on Affine Networks via Lemke's Algorithm," Discussion Paper No. 314. Department of Operations Research, Stanford University (1978).

[5] Beale, E.M.L. and J.A. Tomlin, "Special Facilities in a General Mathematical Programming System for Nonconvex Problems Using Ordered Sets of Variables," in *Proceedings of the Fifth International Conference on Operations Research* (J. Lawrence, ed.), 447-454, Tavistock Publications, London (1970).

[6] Beckman, M.J., "On the Theory of Traffic Flow in Networks," Traffic Quarterly *21*, 109-116 (1967).

[7] Bracken, J. and J.E. Falk, "Computation of Particular Economic Equilibria," IDA Paper P-1208. Institute for Defense Analyses, Arlington, Va. (1976).

[8] Eaves, B.C., "The Linear Complementarity Problem," Management Science *17*, 612-34 (1971).

[9] Eaves, B.C. and R. Saigal, "Homotopies for Computation of Fixed Points on Unbounded Regions," Mathematical Programming *3 (2)*, 225-237 (1972).

[10] Elken, T., The Computation of Economic Equilibria by Path Methods," SOL Technical Report 77-26, Department of Operations Research, Stanford University (1977).

[11] Falk, J.E., "Lagrange Multipliers and Nonconvex Programs," SIAM Journal on Control *7*, 534-545 (1969).

[12] Falk, J.E., "An Algorithm for Locating Approximate Global Solutions of Nonconvex, Separable Problems," Technical Paper Serial T-262, Program in Logistics, The George Washington University (1972).

[13] Falk, J.E. and R.M. Soland, "An Algorithm for Separable Nonconvex Programming Problems," Management Science *15*, 550-569 (1969).

[14] Florian, M. and S. Nguyen, "A Method for Computing Network Equilibrium with Elastic Demands," Transportation Science *8 (4)*, 321-332 (1974).

[15] Grotte, J.H., "Program MOGG—A Code for Solving Separable Nonconvex Optimization Problems," IDA Paper 1318, The Institute for Defense Analyses, Arlington, Va. (1976).

[16] Kuhn, H.W., "On a Theorem of Wald," in *Linear Inequalities and Related Systems* (H.W. Kuhn and A.W. Tucker, eds.), 265-273, Princeton University Press (1956).

[17] Lemke, C.E., "Bimatrix Equilibrium Points and Mathematical Programming," Management Science *11*, 681-89 (1965).

[18] Mangasarian, O.L., "Characterization of Linear Complementarity Problems as Linear Programs," Computer Sciences Technical Report No. 271, Computer Science Department, University of Wisconsin — Madison (1976).

[19] Miller, C.E., "The Simplex Method for Local Separable Programming," in *Recent Advances in Mathematical Programming* (R.L. Graves and P. Wolfe, eds.), 80-100, McGraw Hill, New York (1963).

[20] Nash, J.F., "Equilibrium Points in n-Person Games," Proceedings of the National Academy of Sciences of the U.S.A. *36*, 48-49 (1950).

[21] Negishi, T., "Monopolistic Competition and General Equilibrium," Review of Economic Studies *28 (3)*, 196-201 (1962).

[22] Quirk, J.R. and R. Saposnik, *Introduction to Equilibrium Theory and Welfare Economics*, McGraw-Hill, New York (1968).

[23] Scarf, H., "The Approximation of Fixed Points of a Continuous Mapping," SIAM Journal of Applied Mathematics *15*, 1328-43 (1967).

[24] Scarf, H., *The Computation of Economic Equilibria*, Yale University Press, New Haven (1973).

[25] Soland, R.M., "An Algorithm for Separable Nonconvex Programming Problems II: Nonconvex Constraints," Management Science *17*, 759-773 (1971).

[26] Takayama, T. and G.G. Judge, *Spatial and Temporal Price and Allocation Models*, North-Holland Publishing Company, Amsterdam (1971).

[27] Walras, L., *Elements d'Economie Politique Pure*, Lausanne: Corbaz (1874) (translated as *Elements of Pure Economics*, W. Jaffe, translator, London: Allen and Unwin, 1954).

[28] Wilson, R., "The Bilinear Complementarity Problem and Competitive Equilibria of Linear Economic Models," Econometrica *46 (1)*, (1978).

STOCHASTIC LINEAR PROGRAMS WITH SIMPLE RECOURSE: THE EQUIVALENT DETERMINISTIC CONVEX PROGRAM FOR THE NORMAL, EXPONENTIAL, AND ERLANG CASES

Behram J. Hansotia

*Caterpillar Tractor Co.**
Peoria, Illinois

ABSTRACT

We consider here stochastic linear programs with simple recourse when all the elements of the technology matrix and the resource vector have certain specific distributions. The distributions considered are the Normal, Exponential and Erlang. For the first two instances we extend the equivalent deterministic program to include the variance of the recourse. Finally, a simple example is given to illustrate the application of the formulas for the Erlang case.

1. INTRODUCTION

Recently, Stancu-Minasian and Wets [11] published a bibliography of well over 700 articles dealing with various aspects of the theory and applications of stochastic programming. The applications listed were from such diverse areas as economic planning [12,13] to water storage [7,9] to more classical areas like production planning [3,10] and inventory [4,8]. Needless to say, stochastic programming has flourished considerably in the last two decades and considerable advances have been made both in theory and applications.

In an earlier paper [6] we provided formulas for some special cases of stochastic programs with simple recourse. In this paper we continue along the same lines and present some more expressions for the Exponential, Erlang and Normal cases. The motivation for this research is to provide deterministic equivalent nonlinear programs for a variety of cases, so that researchers can use the nonlinear programs directly after appropriate parameter estimation. The Normal and the Erlang cases are particularly interesting since a large number of unimodal distributions can be approximated by these two distributions.

2. STOCHASTIC PROGRAM WITH SIMPLE RECOURSE

We consider the following stochastic program:

$$\min_x Z = cx + E\,Q(x,\underset{\sim}{\xi})$$

such that

$$Q(x,\underset{\sim}{\xi}) = q_1 y^+ + q_2 y^-,$$

*The paper was written while the author was Associate Professor of Management Science, Bradley University, Peoria, Illinois.

$$y^+ - y^- = \underset{\sim}{T}x - \underset{\sim}{p}\ ,$$

$$Ax = b,\ x \geqslant 0,\ y^+ \geqslant 0,\ y^- \geqslant 0,\ \underset{\sim}{\xi} = (\underset{\sim}{T},\underset{\sim}{p}),$$

where A, b, and c are $(m_1 \times n)$, $(m_1 \times 1)$, and $(1 \times n)$ fixed arrays, x is the $(n \times 1)$ decision vector, and q_1 and q_2 are the $(1 \times m)$ unit penalty vectors. We assume that $\underset{\sim}{\xi} = (\underset{\sim}{p}\ \underset{\sim}{T})$ and $\underset{\sim}{p}$ and $\underset{\sim}{T}$ are random $(m \times 1)$ and $(m \times n)$ arrays. Our objective here is to obtain explicit formulas for Z under specific assumptions about the random arrays $\underset{\sim}{p}$ and $\underset{\sim}{T}$.

A number of authors have studied this problem and it is well known that if $q_1 + q_2 > 0$ the problem is bounded and Z is convex in x (see Williams [15] and Beale [1]).

The problem may be interpreted in two stages. At the first stage we determine an $x \in K = \{x | Ax = b, x \geqslant 0\}$ and at the second stage, after we observe $\underset{\sim}{T}$ and $\underset{\sim}{p}$, a recourse (y^+ *or* y^-) is uniquely determined. The objective here is to select an $x \in K$ which minimizes the sum of the cost of x at the first stage and the expected cost of the recourse at the second stage. $Q(x, \underset{\sim}{\xi})$ is the random variable representing the cost of the recourse under policy x, and the expectation is taken with respect to $\underset{\sim}{\xi}$.

Ziemba [16], Beale [1], and Wets [14] give explicit formulas for Z but only when $\underset{\sim}{p}$ is random. In [6] we give formulas for Z when the elements of $\underset{\sim}{T}$ and $\underset{\sim}{p}$ are independent exponentials and laterally shifted exponentials. In the same paper we also consider the case where the elements of $\underset{\sim}{p}$ and $\underset{\sim}{T}$ are independent Erlangs with shape parameters equal to two. A formula for the distribution of V where $V = \underset{\sim}{c}x + Q(x, \underset{\sim}{\xi})$ is also given for the "independent exponentials" case.

In this paper in Section 3 we consider the case where the elements of $\underset{\sim}{T}$ and $\underset{\sim}{p}$ have a multivariate Normal distribution in $R^{m(n+1)}$ and derive expressions for the expected value and variance of $Q(x, \underset{\sim}{\xi})$. In Section 4 we develop these expressions for the case where the elements of $\underset{\sim}{T}$ and $\underset{\sim}{p}$ are independent Exponentials. Finally, in Section 5 we present recursive equations to compute the expected value of $Q(x, \underset{\sim}{\xi})$ when the elements are independent Erlangs.

3. MULTIVARIATE NORMAL DISTRIBUTION

We assume that $\underset{\sim}{\xi} = (\underset{\sim}{p}, \underset{\sim}{T})$ has a multivariate Normal distribution in $R^{m(n+1)}$. $Q(x, \underset{\sim}{\xi})$ may be written explicitly as

$$Q(x, \underset{\sim}{\xi}) = q_1(\underset{\sim}{T}x - \underset{\sim}{p})^+ + q_2(\underset{\sim}{T}x - \underset{\sim}{p})^- \tag{1}$$

where

$$(\cdot)^+ = \mathrm{Max}(\cdot, 0)$$

and

$$(\cdot)^- = -\mathrm{Min}(\cdot, 0).$$

Denoting the ith element of $\underset{\sim}{T}x - \underset{\sim}{p}$ by $\underset{\sim}{Y}_i$ we have

$$\underset{\sim}{Y}_i = \sum_{j=1}^{n} \underset{\sim}{t}_{ij} x_j - \underset{\sim}{p}_i \tag{2}$$

where $\underset{\sim}{t}_{ij}$ and $\underset{\sim}{p}_i$ are the random elements of the arrays $\underset{\sim}{T}$ and $\underset{\sim}{p}$, respectively. Defining $\mu_{\underset{\sim}{Y}_i}$ and $\sigma^2_{\underset{\sim}{Y}_i}$ as the expected value and variance of Y_i, we have

$$\mu_{\underset{\sim}{Y}_i} = \sum_j x_j E[\underset{\sim}{t}_{ij}] - E[\underset{\sim}{p}_i] \tag{3}$$

and

$$\sigma^2_{\underset{\sim}{Y}_i} = \mathrm{Var}[\underset{\sim}{p}_i] + \sum_j x_j^2 \mathrm{Var}[\underset{\sim}{t}_{ij}] - 2\sum_j x_j \mathrm{Cov}[\underset{\sim}{p}_i, \underset{\sim}{t}_{ij}] + \sum_j \sum_{k \neq j} x_j x_k \mathrm{Cov}[\underset{\sim}{t}_{ij}, \underset{\sim}{t}_{ik}] \tag{4}$$

where E [ ], Var [ ], and Cov [ ] denotes expected value, variance, and covariance, respectively.

Denoting the ratio of mean to standard deviation of Y_i by α_i, i.e.,

$$\alpha_i = \mu_{Y_i}/\sigma_{Y_i}, \tag{5}$$

it is easy to show (see Appendix A, Lemma 1) that

$$E[\underset{\sim}{Y}_i^+] = \mu_{\underset{\sim}{Y}_i}\Phi(\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i), \tag{6}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ represent the cumulative distribution function and the density function of the standard Normal random variable. Similarly,

$$E[\underset{\sim}{Y}_i^-] = -\mu_{\underset{\sim}{Y}_i}\Phi(-\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i). \tag{7}$$

Hence, the equivalent deterministic convex program may be written as

$$\begin{aligned} \min_{x \in K} Z = cx &+ \sum_{i=1}^{m} q_{1i}[\mu_{\underset{\sim}{Y}_i}\Phi(\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i)] + \\ &\sum_{i=1}^{m} q_{2i}[\mu_{\underset{\sim}{Y}_i}\Phi(-\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i)], \end{aligned} \tag{8}$$

$$K = \{x \mid Ax = b,\ x \geqslant 0\}.$$

LEMMA: If $q_1 = q_2 = q$ problem (8) reduces to

$$\min_{x \in K} Z = cx + \sum_{i=1}^{m} q_i[2\mu_{\underset{\sim}{Y}_i}\Phi(\alpha_i) + 2\sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i) - \mu_{\underset{\sim}{Y}_i}] \tag{9}$$

PROOF: Noting that $\Phi(-\alpha_i) = 1 - \Phi(\alpha_i)$, we substitute $q_{1i} = q_{2i} = q_i$ in Eq. (8). For individuals with quadratic utility functions, it can be shown that their expected utility is a linear combination of the expected return and variance of return. Though a quadratic utility function implies increasing risk aversion, this is a first attempt to explicitly incorporate risk in the context of stochastic programming. In that spirit we suggest the following stochastic program:

$$\min_{x \in K} Z^1 = cx + E[Q(x, \underset{\sim}{\xi})] + \lambda \mathrm{Var}\,[Q(x, \underset{\sim}{\xi})], \tag{10}$$

where λ may be interpreted as the decisionmaker's risk-aversion factor.

$$\mathrm{Var}[Q(x, \underset{\sim}{\xi})] = \mathrm{Var}\left[\sum_{i=1}^{m} q_{1i}\underset{\sim}{Y}_i^+ + \sum_{i=1}^{m} q_{2i}\underset{\sim}{Y}_i^-\right] \tag{11}$$

$$= \sum_i q_{1i}^2 \mathrm{Var}[\underset{\sim}{Y}_i^+] + \sum q_{2i}^2 \mathrm{Var}[\underset{\sim}{Y}_i^-] +$$

$$\sum_i \sum_{j \neq i} q_{1i} q_{1j} \mathrm{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^+] +$$

$$\sum_i \sum_{j \neq i} q_{2i} q_{2j} \mathrm{Cov}[\underset{\sim}{Y}_i^-, \underset{\sim}{Y}_j^-] +$$

$$2 \sum_i \sum_j q_{1i} q_{2j} \mathrm{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^-],$$

where $\mathrm{Var}[\underset{\sim}{Y}_i^+]$ can be written as (see Appendix)

(12) $$\mathrm{Var}[\underset{\sim}{Y}_i^+] = \mu^2_{\underset{\sim}{Y}_i} [\Phi(\alpha_i) + \Phi^2(\alpha_i)] +$$
$$\sigma^2_{\underset{\sim}{Y}_i} [\phi^2(\alpha_i) - \alpha_i \phi(\alpha_i) + \Phi(\alpha_i)] +$$
$$2\mu_{\underset{\sim}{Y}_i} \sigma_{\underset{\sim}{Y}_i} [\phi(\alpha_i) + \phi(\alpha_i)\Phi(\alpha_i)].$$

Similarly,

(13) $$\mathrm{Var}[\underset{\sim}{Y}_i^-] = \mu^2_{\underset{\sim}{Y}_i} [\Phi(-\alpha_i) + \Phi^2(-\alpha_i)] +$$
$$\sigma^2_{\underset{\sim}{Y}_i} [\phi^2(\alpha_i) + \alpha_i \phi(\alpha_i) + \Phi(-\alpha_i)] -$$
$$2\mu_{\underset{\sim}{Y}_i} \sigma_{\underset{\sim}{Y}_i} [\phi(\alpha_i) + \phi(\alpha_i)\Phi(-\alpha_i)]$$

We next focus on the covariance terms in expression (11). These are all double integrals and though we cannot reduce them to simple closed-form expressions, we give below their equivalent forms as single integrals. This should considerably reduce the computational effort for evaluating them.

Consider $\mathrm{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^+]$; it may be written as

$$\mathrm{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^+] = E[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^+] - E[\underset{\sim}{Y}_i^+] E[\underset{\sim}{Y}_j^+].$$

Using Jacobians, it can be shown that the joint distribution of $(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j)$ is bivariate Normal. In the Appendix we provide an equivalent form for $E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_j^+]$. This is reproduced below.

(14) $$E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_j^+] = \int_0^\infty y_j [m(y_j)\Phi(\alpha(y_j)) + s(y_j)\phi(\alpha(y_j))] dF(y_j),$$

where

(15) $$\alpha(y_j) = m(y_j)/s(y_j)$$
$$m(y_j) = E[\underset{\sim}{Y}_i | \underset{\sim}{Y}_j = y_j]$$

(16) $$= \mu_{\underset{\sim}{Y}_i} - \rho_{ij} \frac{\sigma_{\underset{\sim}{Y}_i}}{\sigma_{\underset{\sim}{Y}_j}} (y_j - \mu_{\underset{\sim}{Y}_i})$$

(17) $$s(y_j) = \{\mathrm{Var}[\underset{\sim}{Y}_i | \underset{\sim}{Y}_j = y_j]\}^{\frac{1}{2}}$$
$$= \sigma_{\underset{\sim}{Y}_i} (1 - \rho_{ij}^2)$$

where ρ_{ij} is the correlation between $\underset{\sim}{Y}_i$ and $\underset{\sim}{Y}_j$ and is given by (see Lemma 4 in Appendix A)

(18) $$\rho_{ij} = \mathrm{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j)/\sigma_{\underset{\sim}{Y}_i} \sigma_{\underset{\sim}{Y}_j}$$
$$= \frac{1}{\sigma_{\underset{\sim}{Y}_i} \sigma_{\underset{\sim}{Y}_j}} \Big[\sum_k \sum_r x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{jr}) -$$

$$\sum_k x_k\{\mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_i) + \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_j)\} + \mathrm{Cov}(\underset{\sim}{p}_i, \underset{\sim}{p}_j)\Big].$$

The covariance terms in Eq. (18) are obtained directly from the variance-covariance matrix of the multivariate Normal distribution. Similarly,

$$\mathrm{Cov}[\underset{\sim}{Y}_i^-, \underset{\sim}{Y}_j^-] = E[\underset{\sim}{Y}_i^- \underset{\sim}{Y}_j^-] - E[\underset{\sim}{Y}_i^-]E[\underset{\sim}{Y}_j^-]$$

where $E[\underset{\sim}{Y}_i^- \underset{\sim}{Y}_j^-]$ is given by (see Lemma 3 in Appendix A) the following expression:

$$E[\underset{\sim}{Y}_i^- \underset{\sim}{Y}_j^-] = \int_{-\infty}^{0} y_j[m(y_j)\Phi(-\alpha(y_j)) - s(y_j)\,\phi(\alpha(y_j))]dF(y_j). \tag{19}$$

Finally, we need a formula for $\mathrm{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^-]$ in expression (11) to evaluate $\mathrm{Var}(Q(x, \underset{\sim}{\xi}))$:

$$\mathrm{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^-] = E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_j^-] - E[\underset{\sim}{Y}_i^+]E[\underset{\sim}{Y}_j^-]$$

where $E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_j^-]$ is simplified in Lemma 3 in Appendix A and is given by

$$E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_j^-] = \int_0^{\infty} -y_j[m(y_j)\Phi(\alpha(y_j)) + s(y_j)\,\phi(\alpha(y_j))]dF(y_j). \tag{20}$$

Hence for any given policy $x \in K$, expression (11) can be computed if $\underset{\sim}{\xi} = (\underset{\sim}{p}, \underset{\sim}{T})$ is multivariate Normal in $R^{m(n+1)}$.

If the elements of $\underset{\sim}{\xi}$ are correlated, the covariance terms in expression (11) will have to be evaluated numerically; however, we need to perform only single integrals as discussed above.

4. EXPONENTIAL DISTRIBUTION

We consider next the instance where the elements of $\underset{\sim}{\xi} = (\underset{\sim}{p}, \underset{\sim}{T})$ are independent exponentials. In [6] we gave expressions for $E[Q(x, \underset{\sim}{\xi})]$ for this particular case. $E[Q(x, \underset{\sim}{\xi})]$ was obtained by first inverting the bilateral Laplace transform of $\underset{\sim}{Y}_i$ to obtain its density function and then integrating directly. We repeat some of the results here again for completeness and also present formulas for $\mathrm{Var}\, Q(x, \underset{\sim}{\xi})$ so that the mathematical program (10) may be used.

Denoting $1/\lambda_{ij}$ and $1/\lambda_i$ as the means of $\underset{\sim}{t}_{ij}$ and $\underset{\sim}{p}_i$ and defining L_{ik} and L_i thus:

$$L_{ik} = \left\{\lambda_i \Big/ \left(\lambda_i + \frac{\lambda_{ik}}{x_k}\right)\right\} \prod_{j \neq k}\left(\frac{\lambda_{ij}}{x_j}\right)\left(\frac{\lambda_{ij}}{x_j} - \frac{\lambda_{ik}}{x_k}\right), \tag{21}$$

$$L_i = \lambda_i \Pi_j \left(\frac{\lambda_{ij}}{x_j}\right) \Big/ (\lambda_i + \lambda_{ij}\, x_j), \tag{22}$$

we have the following expressions for $f(y_i)$ and $E[Q(x, \underset{\sim}{\xi})]$:

$$f(y_i) = L_i \exp(\lambda_i y_i) \qquad \text{for } y_i < 0, \tag{23a}$$

$$= \sum_{k=1}^{n} L_{ik} \exp\left(\frac{-\lambda_{ik}}{x_k} y_i\right) \qquad \text{for } y_i > 0, \tag{23b}$$

$$= (L_i + \textstyle\sum_k L_{ik})/2 \qquad \text{for } y_i = 0, \tag{23c}$$

and

$$E[Q(x, \underset{\sim}{\xi})] = \sum_{i=1}^{m} q_{1i} E[\underset{\sim}{Y}_i^+] + \sum_{i=1}^{m} q_{2i} E[\underset{\sim}{Y}_i^-], \tag{24}$$

where

$$E[\underset{\sim}{Y}_i^+] = \sum_{k=1}^{n} L_{ik}(x_{ik}/\lambda_{ik})^2 \tag{25}$$

(26) $$E[\underset{\sim}{Y}_i^-] = (1/\lambda_i)^2 L_i.$$

We next consider Var $Q(x, \underset{\sim}{\xi})$:

(27) $$\text{Var}[Q(x,\underset{\sim}{\xi})] = \sum_{i=1}^{m} q_{1i}^2 \; Var\; [\underset{\sim}{Y}_i^+] + \sum_{i=1}^{m} q_{2i}^2 \text{Var}[\underset{\sim}{Y}_i^-] + 2\sum_{i=1}^{m} q_{1i} q_{2i} \text{Cov}[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_i^-].$$

Consider $E[(\underset{\sim}{Y}_i^+)^2]$ and $E[(\underset{\sim}{Y}_i^-)^2]$. We have

(28) $$E[(\underset{\sim}{Y}_i^+)^2] = \int_{0_+}^{\infty} y_i^2 f(y_i)\, dy_i$$

which on integration by parts gives

(29) $$E[(\underset{\sim}{Y}_i^+)^2] = 2\sum_{k=1}^{n} L_{ik}(x_k/\lambda_{ik})^3.$$

Similarly, again integrating by parts gives

(30) $$E[(\underset{\sim}{Y}_i^-)^2] = 2L_i/\lambda_i^3.$$

Therefore

(31) $$\text{Var}\,[\underset{\sim}{Y}_i^+] = 2.\sum_{k=1}^{n} L_{ik}(x_k/\lambda_{ik})^3 - \{E[\underset{\sim}{Y}_i^+]\}^2$$

and

(32) $$\text{Var}\,[\underset{\sim}{Y}_i^-] = 2L_i/\lambda_i^3 - L_i^2/\lambda_i^4.$$

Since $E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_i^-] = 0$, we have

(33) $$\text{Cov}[\underset{\sim}{Y}_i^+,\ \underset{\sim}{Y}_i^-] = -E[\underset{\sim}{Y}_i^+]E[\underset{\sim}{Y}_i^-].$$

Hence Eq. (27) can be computed for any policy x.

5. ERLANG DISTRIBUTION

In [6] we presented an expression for $E(Q(x,\underset{\sim}{\xi}))$ when all the random variables were independent Erlangs with shape parameters equal to two. We extend this result here to the case where the shape parameter is any integer. Inverting the Laplace transform, in this case by the method of residues, requires differentiation with respect to the shape parameter. We present a lemma in this section which may be used to obtain $E(Q(x,\underset{\sim}{\xi}))$. Denoting the means of $\underset{\sim}{t}_{ij}$ and $\underset{\sim}{p}_i$ by r_{ij}/λ_{ij} and r_i/λ_i, respectively, and the bilateral Laplace transform of $\underset{\sim}{Y}_i$ by $L_{\underset{\sim}{Y}_i}(s)$, we have

(34) $$L_{\underset{\sim}{Y}_i}(s) = \left\{\Pi_j \left(\frac{\lambda_{ij}}{x_j}\right)^{r_{ij}} \Big/ \left(s + \frac{\lambda_{ij}}{x_j}\right)^{r_{ij}}\right\}\{(\lambda_i)^{r_i}/(\lambda_i - s)^{r_i}\}$$

with

$$\underset{j}{\text{Max}}\left[-\frac{\lambda_{ij}}{x_j}\right] < s < \lambda.$$

For $s > c$ where

$$\left\{\underset{j}{\text{Max}}\left[-\frac{\lambda_{ij}}{x_j}\right] < c < \lambda_i\right\}$$

there is one pole of order r_i, and the residue R_i of $\exp\{sy_i\}L_{\underset{\sim}{Y}_i}(s)$ at $s = \lambda_i$ is given by

(35) $$R_i = \frac{1}{(r_i - 1)!}\frac{d^{r_i-1}}{ds^{r_i-1}}[(s - \lambda_i)^{r_i} L_{\underset{\sim}{Y}_i}(s) \exp\{sy_i\}]_{s=\lambda_i},$$

$$= \frac{1}{(r_{i-1})!}\frac{d^{r_i-1}}{ds^{r_i-1}}\left[(-\lambda_i)^{r_i}\Pi_j\left(\frac{\lambda_{ij}}{x_j}\right)^{r_{ij}}\Big/\left(s+\frac{\lambda_{ij}}{x_j}\right)^{r_{ij}}\exp\{sy_i\}\right]_{s=\lambda_i}. \tag{36}$$

Hence

$$f(y_i) = -R_i \qquad \text{for} \qquad y_i < 0. \tag{37}$$

For $s < c$ there are n poles each of order r_{ij}, $j = 1, \ldots, n$, and the residue of the k^{th} pole (of order r_{ik}) at $s = -\frac{\lambda_{ik}}{x_k}$ is given by

$$R_{ik} = \frac{1}{(r_{ik-1})!}\frac{d^{r_{ik}-1}}{ds^{r_{ik}-1}}\left[\left(s+\frac{\lambda_{ik}}{x_k}\right)^{r_{ik}} L_{\underset{\sim}{Y}_i}(s)\exp\{sy_i\}\right]_{s=\frac{-\lambda_{ik}}{x_k}} \tag{38}$$

$$= \frac{1}{(r_{ik-1})!}\frac{d^{r_{ik}-1}}{ds^{r_{ik}-1}}\left[\left\{\Pi_j\left(\frac{\lambda_{ij}}{x_j}\right)^{r_{ij}}\Big/\Pi_{j\neq k}\left(s+\frac{\lambda_{ij}}{x_j}\right)^{r_{ij}}\right\}\{(\lambda_i)^{r_i}/(\lambda_i - s)^{r_i}\}\exp\{sy_i\}\right]_{s=-\frac{\lambda_{ik}}{x_k}}. \tag{39}$$

Hence,

$$f(y_i) = \sum_{k=1}^{n} R_{ik} \qquad \text{for} \quad y_i > 0 \tag{40}$$

and

$$f(0) = [\lim_{y_i \to 0^+} f(y_i) + \lim_{y_i \to 0^-} f(y_i)]/2. \tag{41}$$

Denoting the constants in Eqs. (36) and (39) by K_i and L_i, respectively, i.e.,

$$K_i = (-\lambda_i)^{r_i}\Pi_j\left(\frac{\lambda_{ij}}{x_j}\right)^{r_{ij}} \tag{42}$$

and

$$L_i = (\lambda_i)^{r_i}\Pi_j\left(\frac{\lambda_{ij}}{x_j}\right)^{r_{ii}}, \tag{43}$$

we need to determine the (r_{i-1}) th- and (r_{ik-1}) th-order derivatives of G_i and H_i, respectively, where

$$G_i = K_i \exp\{sy_i\}\Pi_j\left(s+\frac{\lambda_{ij}}{x_j}\right)^{-r_{ij}} \tag{44}$$

and

$$H_i = L_i \exp\{sy_i\}\Pi_{\substack{j\\ j\neq k}}(s+\lambda_{ij})^{-r_{ij}}(s-\lambda_i)^{-r_i}. \tag{45}$$

We present a theorem which will enable us to systematically determine the higher order derivatives of these two functions.

THEOREM: The mth-order derivative of the product of n differentiable functions is the mth-order binomial expansion of their sum, with the terms in the expansion interpreted as derivatives as shown below; i.e., given n differentiable functions $f_j(s)$, $j = 1, \ldots, n$,

$$\frac{d^m}{ds^m}\left\{\prod_{j=1}^{n} f_j(s)\right\} = \left\{\sum_{j=1}^{n} f_j(s)\right\}^m,$$

where

$$f_j^{\,i}(s) = \frac{d^i}{ds^i}(f_j(s)) \quad \text{for} \quad i = 1, \ldots, m,$$

$$= f_j(s) \quad \text{for} \quad i = 0.$$

PROOF: The proof is straightforward and uses the results for the derivative of a product of functions. For example, for $n = 3$ and $m = 2$ we have

$$\frac{d^2}{ds^2}\{f_1(s) \cdot f_2(s) \cdot f_3(s)\} = (f_1(s) + f_2(s) + f_3(s))^2$$

$$= f_1^2(s)f_2^0(s)f_3^0(s) + f_1^0(s)f_2^2(s)f_3^0(s) + f_1^0(s)f_2^0(s)f_3^2(s) + \tag{46}$$

$$2f_1^1(s)f_2^1(s)f_3^0(s) + 2f_1^1(s)f_2^0(s)f_3^1(s) + 2f_1^0(s)f_2^1(s)f_3^1(s).$$

Note that to determine the mth-order derivative of the product of functions we need all higher derivatives up to and including m. Therefore, if in expression (44) we denote $\exp\{sy_i\}$ by $f_1(s)$ and $\Pi_j\left(s + \frac{\lambda_{ij}}{x_j}\right)^{r_{ij}}$ by $f_2(s)$, we have for any integer $m > 0$

$$\frac{d^m}{ds^m} f_1(s) = y_i^m \exp\{sy_i\}. \tag{47}$$

Denoting $\left(s + \frac{\lambda_{ij}}{x_j}\right)^{-r_{ij}}$ by $g_j(s)$ we have

$$\frac{d^m}{ds^m}(f(s)) = \frac{d^m}{ds^m}\left[\prod_{j=1}^{n}\left(s + \frac{\lambda_{ij}}{x_j}\right)^{-r_{ij}}\right]$$

$$= \frac{d^m}{ds^m}\left[\prod_{j=1}^{n} g_j(s)\right],$$

which according to the above theorem is

$$\left[\sum_{j=1}^{n} g_j(s)\right]^m, \tag{48}$$

which can be expanded in a manner similar to Eq. (46). Note that

$$g_j^0(s) = g_j(s)$$

and

$$g^e(s) = \frac{d^e}{ds^e}(g_j(s))$$

$$= \left[\prod_{u=0}^{e-1}(-r_{ij} - u)\right]\left(s + \frac{\lambda_{ij}}{x_j}\right)^{-r_{ij}-e} \quad \text{for } e = 1, \ldots, m. \tag{49}$$

If $m = r_{i-1}$, we can compute all the higher order derivatives required in expression (36), and $f(y_i)$, $y_i < 0$ can be computed. The partial expectation of $\underset{\sim}{Y}_i$ for $y_i < 0$ can then be computed by direct integration. Similarly, in expression (45) we define $f_1(s) = \exp(sy_i)$, $f_2(s) = \prod_{j \neq k}(s + \lambda_{ij})^{-r_{ij}}$, and $f_3(s) = (s - \lambda_i)^{-r_i}$.

Equations (47) through (49) go through again for the higher order derivatives with a slight modification of expression (48); namely, j cannot equal k in expression (48). Denoting this modified expression as (48[1]) we have

$$\frac{d^m}{ds^m}(f(s)) = \left[\sum_{\substack{j=1 \\ j\neq k}}^{n} g_j(s)\right]^m \tag{48[1]}$$

and

$$\frac{d^m}{ds^m}(f_3(s)) = \frac{d^m}{ds^m}(s-\lambda_i)^{-r_i}$$

$$= \left[\prod_{u=0}^{m-1}(-r_i-u)\right](s-\lambda_i)^{-r_i-m}. \tag{50}$$

Note in Eq. (39) $m = r_{ik-1}$. Therefore R_{ik} and hence $f(y_i)$, $y_i > 0$ can be obtained. The partial expected value of $\underset{\sim}{Y}_i$ for $y_i > 0$ is then obtained by direct integration.

EXAMPLE: We consider a problem where $\underset{\sim}{T}$ is (2×2) with each element of $\underset{\sim}{T}$ and $\underset{\sim}{p}$ having an independent Erlang distribution. Denoting R as the matrix of shape parameters of $\underset{\sim}{T}$ and h as the vector of shape parameters of $\underset{\sim}{p}$ we assume

$$R = \begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \text{ and } h = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

Note that the (i,j)th element of R is r_{ij} and the ith element of h is r_i. Using Eqs. (36) through (49) it can be shown that

$$f(y_1) = \exp\{\lambda_1 y_1\}[L_1(x)y_1^2 + L_2(x)y_1 + L_3(x)] \quad \text{for} \quad y_1 < 0$$

and

$$f(y_i) = \exp\left\{-\frac{\lambda_{12}}{x_2}y_1\right\}[M_1(x)y_1^2 + M_2(x)y_1 + M_3(x)] +$$

$$\exp\left\{-\frac{\lambda_{11}}{x_1}y_1\right\}[M_4(x)y_1 + M_5(x)] \quad \text{for} \quad y_1 > 0.$$

Hence,

$$E[\underset{\sim}{Y}_1^-] = 6\frac{L_1(x)}{\lambda_1^4} - 2\frac{L_2(x)}{\lambda_1^3} + \frac{L_3(x)}{\lambda_1^2}$$

and

$$E[\underset{\sim}{Y}_1^+] = \frac{6M_1(x)}{\left(\frac{\lambda_{12}}{x_2}\right)^4} + \frac{2M_2(x)}{\left(\frac{\lambda_{12}}{x_2}\right)^3} + \frac{M_3(x)}{\left(\frac{\lambda_{12}}{x_2}\right)^2} + \frac{2M_4(x)}{\left(\frac{\lambda_{11}}{x_1}\right)^3} + \frac{M_5(x)}{\left(\frac{\lambda_{11}}{x_1}\right)^2}.$$

Similarly,

$$f(y_2) = \exp\{\lambda_2 y_2\}(P_1(x)y_2 + P_2(x)) \qquad \text{for } y_2 < 0$$

and

$$f(y_2) = \exp\{-\frac{\lambda_{22}}{x_2}y_2\}[N_1(x)y_2^3 + N_2(x)y_2^2 + N_3(x)y_2 + N_4(x)] +$$

$$\exp\{-\frac{\lambda_{21}}{x_1}y_2\}[N_5(x)y_2^2 + N_6(x)y_2 + N_7(x)] \qquad \text{for } y_2 > 0.$$

Therefore,

$$E[\underset{\sim}{Y}_2^-] = -\frac{2P_1(x)}{\lambda_2^3} + \frac{P_2(x)}{\lambda_2^2}$$

and

$$E[\underset{\sim}{Y}_2^+] = \frac{24N_1(x)}{\left(\frac{\lambda_{22}}{x_2}\right)^5} + \frac{6N_2(x)}{\left(\frac{\lambda_{22}}{x_2}\right)^4} + \frac{2N_3(x)}{\left(\frac{\lambda_{22}}{x_2}\right)^3} + \frac{N_4(x)}{\left(\frac{\lambda_{22}}{x_2}\right)^2} + \frac{6N_5(x)}{\left(\frac{\lambda_{21}}{x_1}\right)^4} + \frac{2N_6(x)}{\left(\frac{\lambda_{21}}{x_1}\right)^3} + \frac{N_7(x)}{\left(\frac{\lambda_{21}}{x_1}\right)^2}.$$

The functions $L_i(x)$, $M_j(x)$, $N_k(x)$, and $P_m(x)$ for $i = 1, 2, 3$, $j = 1, \ldots, 5$, $k = 1, \ldots, 7$, and $m = 1, 2$ are reported in Appendix B. Since all the partial expected values are available, $Q(x, \underset{\sim}{\xi})$ can be computed.

4. SUMMARY

We presented here equivalent deterministic convex programs for certain special instances of stochastic programs with simple recourse. The problem studied had *all* the elements of the technology matrix and the recource vector as random variables. Previous studies (except Hansotia [6]) reported explicit formulas for cases where only the resourse vector was considered random. The specific cases studied here had all the random variables generated by a multivariate Normal distribution, independent Exponential distributions, and independent Erlang distributions. For the first two cases we also extended the equivalent deterministic program to include the variance of the recourse. Finally, a simple example was given to illustrate the use of the "not-so-simple" formulas for the Erlang case.

BIBLIOGRAPHY

[1] Beale, E.M.L., "On Minimizing a Convex Function Subject to Linear Inequalities," Journal of the Royal Statistical Society, Series B *17*, 173-184 (1955).

[2] Beale, E.M.L., "The Use of Quadratric Programming in Stochastic Linear Programming," Tech. Report, Rand Corporation, Santa Monica (1961).

[3] Beebe, J.H., C.S. Beightler, and J.P. Stark, "Stochastic Optimization of Production Planning," Operations Research *16*, 799-818 (1968).

[4] El-Agizy, M., "Dynamic Inventory Models and Stochastic Program," IBM Journal of Research and Development *13*, 351-356 (1969).

[5] Hansotia, B., *On Stochastic Programming*, unpublished Ph.D. dissertation, College of Business Administration, University of Illinois, Urbana (1973).

[6] Hansotia, B., "Some Special Cases of Stochastic Programs with Recourse," Operations Research *25*, 361-363 (1977).

[7] Loucks, D.P., "Discrete Chance-Constrained Models for River Basin Planning," in *Stochastic Programming*, Proceedings of the 1974 Oxford International Conference, M. Dempster (ed), Academic Press, New York (1976).

[8] Oral, M., and S.S. Rao, "An Objective Function for Inventory Control Models," in *Stochastic Programming*, Proceedings of the 1974 Oxford International Conference, M. Dempster (ed), Academic Press, New York (1976)

[9] Revelle, C., and J. Gundelach, "A Variance Minimizing Linear Decision Rule for Reservoir Management and Design," in *Stochastic Programming*, Proceedings of the 1974 Oxford International Conference, M. Dempster (ed), Academic Press, New York (1976).

[10] Sengupta, J.K., and J.H. Portillo-Campbell, "A Reliability Programming Approach to Production Planning," International Statistical Review *41*, 115-127 (1973).

[11] Stancu-Minasian, I.M., and M.J. Wets, "A Research Bibliography in Stochastic Programming, 1955-1975," Operations Research *24*, 1078-1119 (1976).

[12] Tintner, G., "The Use of Stochastic Linear Programming in Planning," Indian Economic Review *5*, 159-167 (1960).

[13] Tintner, G., and N.S. Raghaven, "Stochastic Linear Programming Applied to a Dynamic Planning Model in India," Economia Internazaionale *23*, 3-15 (1970).

[14] Wets, R., "Programming Under Uncertainty: The Complete Problem," A. Wahrscheinlichkeitstheorie Verw. Gebiete, *4*, 316-339 (1966).

[15] Williams, A.C., "Approximation Formulas for Stochastic Linear Programming," SIAM Journal *14*, 668-677 (1966).

[16] Ziemba, W., "Stochastic Programs with Simple Recourse," in *Mathematical Programming in Theory and Practice,* pp. 213-274, P. Hammer and G. Zoutendijk (eds), North-Holland, Amsterdam (1974).

APPENDIX A

LEMMA 1: If $\underset{\sim}{Y}_i \sim N(\mu_{\underset{\sim}{Y}_i}, \sigma_{\underset{\sim}{Y}_i})$ and $\mu_{\underset{\sim}{Y}_i}/\sigma_{\underset{\sim}{Y}_i} = \alpha_i$, then

$$E[\underset{\sim}{Y}_i^+] = \mu_{\underset{\sim}{Y}_i}\Phi(\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i)$$

and

$$E[\underset{\sim}{Y}_i^-] = -\mu_{\underset{\sim}{Y}_i}\Phi(-\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and the cumulative distribution functions of the standard Normal random variable.

PROOF:

(A1) $$E[\underset{\sim}{Y}_i^+] = \int_0^\infty \frac{y}{\sqrt{2\pi}\sigma_{\underset{\sim}{Y}_i}} \exp\left\{-\frac{1}{2}[(y - \mu_{\underset{\sim}{Y}_i})/\sigma_{\underset{\sim}{Y}_i}]^2\right\} dy .$$

Substituting $(y - \mu_{\underset{\sim}{Y}_i})/\sigma_{\underset{\sim}{Y}_i} = z$ we have the right-hand side (RHS) of Eq. (A1) equal to

(A2) $$\frac{\sigma_{\underset{\sim}{Y}_i}}{\sqrt{2\pi}} \int_{-\alpha_i}^\infty z \exp\left\{-\frac{1}{2}z^2\right\} dz + \frac{\mu_{\underset{\sim}{Y}_i}}{\sqrt{2\pi}} \int_{-\alpha_i}^\infty \exp\left\{-\frac{1}{2}z^2\right\} dz .$$

The second expression in Eq. (A2) is merely $\mu_{\underset{\sim}{Y}_i}[1 - \Phi(-\alpha_i)]$, which may also be written as $\mu_{\underset{\sim}{Y}_i}\Phi(\alpha_i)$. Substituting v for $\frac{1}{2}z^2$ in the first integral, it becomes

$$\frac{\sigma_{\underset{\sim}{Y}_i}}{\sqrt{2\pi}} \int_{\frac{1}{2}\alpha_i^2}^\infty e^{-v} dv = \sigma_{\underset{\sim}{Y}_i} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\alpha_i^2\right\}$$

$$= \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i) .$$

Hence $E[\underset{\sim}{Y}_i^+]$ is given by

$$\mu_{\underset{\sim}{Y}_i}\Phi(\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i) .$$

To prove the second result we note that

(A3) $$\underset{\sim}{Y}_i^- = [-\underset{\sim}{Y}_i]^+$$

Denoting $-\underset{\sim}{Y}_i$ by $\underset{\sim}{W}$ and the mean and standard deviation of $\underset{\sim}{W}$ by $\mu_{\underset{\sim}{W}}$ and $\sigma_{\underset{\sim}{W}}$, we have

$$E[\underset{\sim}{W}^+] = \mu_{\underset{\sim}{W}}\Phi(\alpha) + \sigma_{\underset{\sim}{W}}\phi(\alpha)$$

where $\alpha = \mu_{\underset{\sim}{W}}/\sigma_{\underset{\sim}{W}}$. Noting that $\mu_{\underset{\sim}{W}} = -\mu_{\underset{\sim}{Y}_i}$ and $\sigma_{\underset{\sim}{W}} = \sigma_{\underset{\sim}{Y}_i}$, we have

$$E[\underset{\sim}{Y}_i^-] = -\mu_{\underset{\sim}{Y}_i}\Phi(-\alpha_i) + \sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i)\,.$$

Q.E.D.

Note how similar $E[\underset{\sim}{Y}_i^+]$ and $E[\underset{\sim}{Y}_i^-]$ are. To obtain the latter all we need to do is replace $\mu_{\underset{\sim}{Y}_i}$ by $-\mu_{\underset{\sim}{Y}_i}$ and $-\alpha_i$ by α_i. Also note that because of symmetry $\phi(-\alpha_i) = \phi(\alpha_i)$.

LEMMA 2: If $\underset{\sim}{Y}_i \sim N(\mu_{\underset{\sim}{Y}_i}, \sigma_{\underset{\sim}{Y}_i})$ and $\alpha_i = \mu_{\underset{\sim}{Y}_i}/\sigma_{\underset{\sim}{Y}_i}$, then,

$$\mathrm{Var}[\underset{\sim}{Y}_i^+] = \mu_{\underset{\sim}{Y}_i}^2[\Phi(\alpha_i) + \Phi^2(\alpha_i)] + \sigma_{\underset{\sim}{Y}_i}^2[\phi^2(\alpha_i) - \alpha_i\phi(\alpha_i) + \Phi(\alpha_i)] + 2\mu_{\underset{\sim}{Y}_i}\sigma_{\underset{\sim}{Y}_i}[\phi(\alpha_i) + \phi(\alpha_i)\Phi(\alpha_i)]$$

and

$$\mathrm{Var}[\underset{\sim}{Y}_i^-] = \mu_{\underset{\sim}{Y}_i}^2[\Phi(-\alpha_i) + \Phi^2(-\alpha_i)] + \sigma_{\underset{\sim}{Y}_i}^2[\phi^2(\alpha_i) + \alpha_i\phi(\alpha_i) + \Phi(-\alpha_i)] - 2\mu_{\underset{\sim}{Y}_i}\sigma_{\underset{\sim}{Y}_i}[\phi(\alpha_i) + \phi(\alpha_i)\Phi(-\alpha_i)]\,.$$

PROOF:

(A4) $$\mathrm{Var}[\underset{\sim}{Y}_i^+] = E[\underset{\sim}{Y}_i^+)^2] - \{E[\underset{\sim}{Y}_i^+]\}^2$$

(A5) $$E[(\underset{\sim}{Y}_i^+)^2] = \int_0^\infty y^2 df_{\underset{\sim}{Y}_i}(y_i)$$

Substituting $z = (y - \mu_{\underset{\sim}{Y}_i})/\sigma_{\underset{\sim}{Y}_i}$ in Eq. (A4), we have the RHS of Eq. (A4) equal to

$$\frac{1}{\sqrt{2\pi}}\left[\sigma_{\underset{\sim}{Y}_i}^2\int_{-\alpha_i}^\infty z^2\exp\left\{-\frac{1}{2}z^2\right\}dz + 2\mu_{\underset{\sim}{Y}_i}\sigma_{\underset{\sim}{Y}_i}\int_{-\alpha_i}^\infty 2\exp\left\{-\frac{1}{2}z^2\right\}dz + \mu_{\underset{\sim}{Y}_i}^2\int_{-\alpha_i}^\infty \exp\left\{-\frac{1}{2}z^2\right\}dz\right].$$

$\frac{1}{\sqrt{2\pi}}\int_{-\alpha_i}^\infty z\exp\left\{-\frac{1}{2}z^2\right\}dz$ is merely $\phi(\alpha_i)$ (see Lemma 1) and

$$\frac{1}{\sqrt{2\pi}}\int_{-\alpha_i}^\infty \exp\left\{-\frac{1}{2}z^2\right\}dz = 1 - \Phi(-\alpha_i) = \Phi(\alpha_i)\,.$$

Hence, let us consider $I = \frac{1}{\sqrt{2\pi}}\int_{-\alpha_i}^\infty z^2\exp\left\{-\frac{1}{2}z^2\right\}dz$. Integrating by parts and noting that

$$\int z\exp\left\{-\frac{1}{2}z^2\right\}dz = -\exp\left\{-\frac{1}{2}z^2\right\} \text{ (up to a constant)},$$

we have I equal to

$$\frac{1}{\sqrt{2\pi}}\left[-z\exp\left(-\frac{1}{2}z^2\right)\Big|_{-\alpha_i}^{\infty}+\exp\int_{-\alpha_i}^{\infty}\left\{-\frac{1}{2}z^2\right\}dz\right]$$

$$=-\alpha_i\phi(\alpha_i)+1-\Phi(-\alpha_i)$$

$$=-\alpha_i\phi(\alpha_i)+\Phi(\alpha_i)\,.$$

Hence $E[(\underset{\sim}{Y}_i^+)^2]$ is given by

$$\sigma_{\underset{\sim}{Y}_i}^2[-\alpha_i\phi(\alpha_i)+\Phi(\alpha_i)]+2\mu_{\underset{\sim}{Y}_i}\sigma_{\underset{\sim}{Y}_i}\phi(\alpha_i)+\mu_{\underset{\sim}{Y}_i}{}^2\Phi(\alpha_i)\,. \qquad \text{Q.E.D.}$$

Substituting $E[\underset{\sim}{Y}_i^+]$ (from Lemma 1) into (A5) and using the above result we have

$$\mathrm{Var}[\underset{\sim}{Y}_i^+]=\mu_{\underset{\sim}{Y}_i}^2[\Phi(\alpha_i)+\Phi^2(\alpha_i)]+\sigma_{\underset{\sim}{Y}_i}{}^2[\phi^2(\alpha_i)-\alpha_i\phi(\alpha_i)+\Phi(\alpha_i)]+$$

(A6) $$2\mu_{\underset{\sim}{Y}_i}\sigma_{\underset{\sim}{Y}_i}[\phi(\alpha_i)+\phi(\alpha_i)\Phi(\alpha_i)]\,.$$

Q.E.D.

The second result is obtained quite easily by substituting into Eq. (A6) $-\mu_{\underset{\sim}{Y}_i}$ for $\mu_{\underset{\sim}{Y}_i}$ and $-\alpha_i$ for α_i and noting that $\phi(-\alpha_i)=\phi(\alpha_i)$.

LEMMA 3: If $(\underset{\sim}{Y}_i,\ \underset{\sim}{Y}_j)$ is Normal in R^2, then

(i) $$E[\underset{\sim}{Y}_i^+\underset{\sim}{Y}_j^+]=\int_0^{\infty}y_j[m(y_j)\Phi(\alpha(y_j))+s(y_j)\phi(\alpha(y_j))]dF(y_j),$$

(ii) $$E[\underset{\sim}{Y}_i^-\underset{\sim}{Y}_j^-]=\int_{-\infty}^{0}y_j[m(y_j)\Phi(-\alpha(y_j))-s(y_j)\phi(\alpha(y_j))]dF(y_j)\,,$$

and

(iii) $$E[\underset{\sim}{Y}_i^+\underset{\sim}{Y}_j^-]=\int_0^{\infty}-y_j[m(y_j)\Phi(\alpha(y_j))+s(y_j)\phi(\alpha(y_j))]dF(y_j)\,,$$

where

where $m(y_j)=E[\underset{\sim}{Y}_i|\underset{\sim}{Y}_j=y_j]$,

$s(y_j)=\{\mathrm{Var}[\underset{\sim}{Y}_i|\underset{\sim}{Y}_j=y_j]\}^{\frac{1}{2}}$, and

$\alpha(y_j)=m(y_j)/s(y_j)\,.$

PROOF:

(i) $$E\,[\underset{\sim}{Y}_i^+\,\underset{\sim}{Y}_j^+]=\int_0^{\infty}\int_0^{\infty}y_iy_j\,dF(y_i,y_j)$$

(A7) $$=\int_0^{\infty}\int_0^{\infty}y_iy_j\,dF\,(y_i|y_j)\,dF\,(y_j)\,.$$

But $dF\,(y_i|y_j)$ is Normal in R with mean $m\,(y_j)$ and standard deviation $s\,(y_j)$, where

(A8) $$m(y_j)=\mu_{\underset{\sim}{Y}_i}-\rho_{ij}\frac{\sigma_{\underset{\sim}{Y}_i}}{\sigma_{\underset{\sim}{Y}_j}}(y_j-\mu_{\underset{\sim}{Y}_i})$$

and

$$s(y_j) = \sigma_{\underset{\sim}{Y}_i}(1 - \rho_{ij}^2) \tag{A9}$$

with ρ_{ij} being the correlation between $\underset{\sim}{Y}_i$ and $\underset{\sim}{Y}_j$. Equation (A7) may be written as

$$\int_0^\infty y_j \left[\int_0^\infty y_i \, dF(y_i|y_j) \right] dF(y_j) = \int_0^\infty y_j \, E[Y_i^+ | \underset{\sim}{Y}_j = y_j] \, dF(y_j) .$$

Using Lemma 1, we have

$$E[\underset{\sim}{Y}_i^+ | \underset{\sim}{Y}_j = y_j] = m(y_j)\Phi(\alpha(y_j)) + s(y_j)\phi(\alpha(y_j)) ,$$

where

$$\alpha(y_j) = m(y_j)/s(y_j) . \tag{A10}$$

Hence $E[\underset{\sim}{Y}_i^+ \underset{\sim}{Y}_j^+]$ is given by

$$\int_0^\infty y_j [m(y_j)\Phi(\alpha(y_j) + s(y_j)\phi(\alpha(y_j))] dF(y_j) .$$

(ii) In an identical fashion we prove the second result:

$$E[\underset{\sim}{Y}_i^- \underset{\sim}{Y}_j^-] = \int_{-\infty}^0 \int_{-\infty}^0 y_i y_j \, dF(y_i|y_j) \, dF(y_j) = \int_{-\infty}^0 - y_j \, E[\underset{\sim}{Y}_i^- | \underset{\sim}{Y}_j = y_j] \, dF(y_j) .$$

Using Lemma 1 we have

$$E[\underset{\sim}{Y}_i^- | \underset{\sim}{Y}_j = y_j] = -m(y_j)\Phi(-\alpha(y_j)) + s(y_j)\phi(\alpha(y_j)$$

where $m(y_j) = E[\underset{\sim}{Y}_i | \underset{\sim}{Y}_j = y_j]$ and is given by expression (A8). $s(y_j) = \{\text{Var}[\underset{\sim}{Y}_i | \underset{\sim}{Y}_j = y_j]\}^{1/2}$ and is given by expression (A9) and $\alpha(y_j) = m(y_j)/s(y_j)$. Hence

$$E[\underset{\sim}{Y}_i^- \underset{\sim}{Y}_j^-] = \int_{-\infty}^0 y_j [m(y_j)\Phi(-\alpha(y_j)) - s(y_j)\phi(\alpha(y_j))] dF(y_j)$$

Q.E.D.

Finally,

$$\begin{aligned} E[\underset{\sim}{Y}_i^+, \underset{\sim}{Y}_j^-] &= \int_{-\infty}^0 \int_0^\infty - y_j y_i dF(y_i, y_j) \\ &= \int_{-\infty}^0 - y_j \left[\int_0^\infty y_i dF(y_i|y_j) \right] dF(y_j) \\ &= \int_{-\infty}^0 - y_j [m(y_j)\Phi(\alpha(y_j)) + s(y_j)\phi(\alpha(y_j))] \, dF(y_j) \end{aligned}$$

Q.E.D.

Note that $\rho_{ij} = \text{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j)/\sigma_{\underset{\sim}{Y}_i} \sigma_{\underset{\sim}{Y}_j}$; we present next a Lemma to compute $\text{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j)$.

LEMMA 4: If $\underset{\sim}{Y}_i = \Sigma_k \underset{\sim}{t}_{ik} x_k - \underset{\sim}{p}_i$ and $\underset{\sim}{Y}_j = \Sigma_k \underset{\sim}{t}_{jk} x_k - \underset{\sim}{p}_j$,

$$\begin{aligned} \text{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j) = {} & \Sigma_k \Sigma_r x_k x_r \text{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{jr}) - \Sigma_k x_k \text{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_i) - \\ & \Sigma_k x_k \text{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_j) + \text{Cov}(\underset{\sim}{p}_i, \underset{\sim}{p}_j) . \end{aligned}$$

PROOF: Consider

$$\text{Var}[\underset{\sim}{Y}_i + \underset{\sim}{Y}_j] = \text{Var}[\underset{\sim}{Y}_i] + \text{Var}[\underset{\sim}{Y}_j] + 2\text{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j) \tag{A11}$$

and

$$\text{Var}[\underset{\sim}{Y}_i - \underset{\sim}{Y}_j] = \text{Var}[\underset{\sim}{Y}_i] + \text{Var}[\underset{\sim}{Y}_j] - 2\text{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j) . \tag{A12}$$

Taking the difference between expression (A11) and (A12), we have

$$\text{(A13)} \qquad \mathrm{Var}[\underset{\sim}{Y}_i + \underset{\sim}{Y}_j] - \mathrm{Var}[\underset{\sim}{Y}_i - \underset{\sim}{Y}_j] = 4\mathrm{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j).$$

But,

$$\begin{aligned}\mathrm{Var}[\underset{\sim}{Y}_i + \underset{\sim}{Y}_j] = {} & \Sigma_k x^2(\mathrm{Var}(\underset{\sim}{t}_{ik}) + \mathrm{Var}(\underset{\sim}{t}_{jk})) + \mathrm{Var}(\underset{\sim}{p}_i) + \mathrm{Var}(\underset{\sim}{p}_j) + \\ & \Sigma_k \sum_{r\neq k} x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{ir}) - 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_i) + \\ & \Sigma_k \sum_{r\neq k} x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{t}_{jr}) - 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_j) + \\ & 2\Sigma_k \Sigma_r x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{jr}) - 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_i) - \\ & 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_j) + 2\mathrm{Cov}(\underset{\sim}{p}_i, \underset{\sim}{p}_j).\end{aligned} \qquad \text{(A14)}$$

Similarly,

$$\begin{aligned}\mathrm{Var}[\underset{\sim}{Y}_i - \underset{\sim}{Y}_j] = {} & \Sigma_k x_k^2(\mathrm{Var}(\underset{\sim}{t}_{ik}) - \mathrm{Var}(\underset{\sim}{t}_{jk})) + \mathrm{Var}(\underset{\sim}{p}_i) + \mathrm{Var}(\underset{\sim}{p}_j) + \\ & \Sigma_k \sum_{r\neq k} x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{ir}) - 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_i) + \\ & \Sigma_k \sum_{r\neq k} x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{t}_{jr}) - 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_j) - \\ & 2\Sigma_k \Sigma_r x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{jr}) + 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_i) + \\ & 2\Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_j) - 2\mathrm{Cov}(\underset{\sim}{p}_i, \underset{\sim}{p}_j).\end{aligned} \qquad \text{(A15)}$$

Substituting expressions (A14) and (A15) into expression (A13) and simplifying, we have

$$\mathrm{Cov}(\underset{\sim}{Y}_i, \underset{\sim}{Y}_j) = \Sigma_k \Sigma_r x_k x_r \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{t}_{jr}) - \Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{jk}, \underset{\sim}{p}_i) - \Sigma_k x_k \mathrm{Cov}(\underset{\sim}{t}_{ik}, \underset{\sim}{p}_j) + \mathrm{Cov}(\underset{\sim}{p}_i, \underset{\sim}{p}_j).$$

Q.E.D.

APPENDIX B

The following functions appear in the example in Section 3.

1. $L_1(x) = -\dfrac{K_1}{2!}\left[\dfrac{\lambda_{11}}{x_1} + \lambda_1\right]^{-2}\left[\dfrac{\lambda_{12}}{x_2} + \lambda_1\right]^{-3}$

2. $L_2(x) = -K_1\left[3\left[\dfrac{\lambda_{11}}{x_1} + \lambda_1\right]^{-2}\left[\dfrac{\lambda_{12}}{x_2} + \lambda_1\right]^{-4} - 2\left[\dfrac{\lambda_{11}}{x_1} + \lambda_1\right]^{-3}\left[\dfrac{\lambda_{12}}{x_2} + \lambda_1\right]^{-3}\right]$

3. $L_3(x) = -\dfrac{K_1}{2!}\left[6\left[\dfrac{\lambda_{11}}{x_1} + \lambda_1\right]^{-4}\left[\dfrac{\lambda_{12}}{x_2} + \lambda_1\right]^{-3} + 12\left[\dfrac{\lambda_{11}}{x_1} + \lambda_1\right]^{-3}\left[\dfrac{\lambda_{12}}{x_2} + \lambda_1\right]^{-4} + 12\left[\dfrac{\lambda_{11}}{x_1} + \lambda_1\right]^{-2}\left[\dfrac{\lambda_{12}}{x_2} + \lambda_1\right]^{-5}\right]$

4. $M_1(x) = \dfrac{K_2}{2!}\left[\dfrac{\lambda_{11}}{x_1} - \dfrac{\lambda_{12}}{x_2}\right]^{-2}\left[-\dfrac{\lambda_{12}}{x_2} - \lambda_1\right]^{-3}$

5. $M_2(x) = -\dfrac{K_2}{2!}\left[2\left[\dfrac{\lambda_{11}}{x_1} - \dfrac{\lambda_{12}}{x_2}\right]^{-3}\left[-\dfrac{\lambda_{12}}{x_2} - \lambda_1\right]^{-3} + 3\left[\dfrac{\lambda_{11}}{x_1} - \dfrac{\lambda_{12}}{x_2}\right]^{-2}\left[-\dfrac{\lambda_{12}}{x_2} - \lambda_1\right]^{-4}\right]$

6. $$M_3(x) = 3K_2\left[\left(\frac{\lambda_{11}}{x_1} - \frac{\lambda_{12}}{x_2}\right)^{-4}\left(-\frac{\lambda_{12}}{x_2} - \lambda_1\right)^{-3} + 2\left(\frac{\lambda_{11}}{x_1} - \frac{\lambda_{12}}{x_2}\right)^{-2}\left(-\frac{\lambda_{12}}{x_2} - \lambda_1\right)^{-3}\right]$$

7. $$M_4(x) = K_1\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{11}}{x_1}\right)^{-3}\left(-\frac{\lambda_{11}}{x_1} - \lambda_1\right)^{-3}$$

8. $$M_5(x) = -3K_1\left[\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{11}}{x_1}\right)^{-4}\left(-\frac{\lambda_{11}}{x_1} - \lambda_1\right)^{-3} + \left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{11}}{x_1}\right)^{-3}\left(-\frac{\lambda_{11}}{x_1} - \lambda_1\right)^{-4}\right]$$

9. $$P_1(x) = K_2\Pi_j\left(\frac{\lambda_{2j}}{x_j} + \lambda_2\right)^{-r_{2j}}$$

10. $$P_2(x) = \left[-4\left(\frac{\lambda_{21}}{x_1} + \lambda_2\right)^{-3}\left(\frac{\lambda_{22}}{x_2} + \lambda_2\right)^{-5} - 3\left(\frac{\lambda_{21}}{x_1} + \lambda_2\right)^{-4}\left(\frac{\lambda_{22}}{x_2} + \lambda_2\right)^{-4}\right]$$

11. $$N_1(x) = \frac{K_2}{3!}\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-3}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-2}$$

12. $$N_2(x) = -\frac{3K_2}{3!}\left\{3\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-4}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-2} + 2\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-3}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-3}\right\}$$

13. $$N_3(x) = \frac{K_2}{2!}\left\{12\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-5}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-2} + 6\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-3}\cdot\right.$$
$$\left.\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-4} + 12\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-4}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-3}\right\}$$

14. $$N_4(x) = \frac{K_2}{3!}\left\{-60\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-6}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-2} - 24\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-3}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-5} -\right.$$
$$\left.72\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-5}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-3} - 54\left(\frac{\lambda_{21}}{x_1} - \frac{\lambda_{22}}{x_2}\right)^{-4}\left(-\frac{\lambda_{22}}{x_2} - \lambda_2\right)^{-4}\right\}$$

15. $$N_5(x) = \frac{K_2}{2!}\left\{\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{21}}{x_1}\right)^{-3}\left(-\frac{\lambda_{21}}{x_1} - \lambda_2\right)^{-2}\right\}$$

16. $$N_6(x) = \frac{K_2}{2!}\left\{-6\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{21}}{x_1}\right)^{-4}\left(-\frac{\lambda_{21}}{x_1} - \lambda_2\right)^{-2} - 4\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{21}}{x_1}\right)^{-3}\left(-\frac{\lambda_{21}}{x_1} - \lambda_2\right)^{-3}\right\}$$

17. $$N_7(x) = \frac{K_2}{2!}\left\{12\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{21}}{x_1}\right)^{-5}\left(-\frac{\lambda_{21}}{x_1} - \lambda_2\right)^{-2} + 6\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{21}}{x_1}\right)^{-3}\cdot\right.$$
$$\left.\left(-\frac{\lambda_{21}}{x_1} - \lambda_2\right)^{-4} + 12\left(\frac{\lambda_{12}}{x_2} - \frac{\lambda_{21}}{x_1}\right)^{-4}\left(-\frac{\lambda_{21}}{x_1} - \lambda_2\right)^{-3}\right\}$$

PARTIALLY CONTROLLED DEMAND AND INVENTORY CONTROL: AN ADDITIVE MODEL

Yves Balcer

Department of Economics
University of Wisconsin-Madison
Madison, Wisconsin

ABSTRACT

The primary goal of this paper is to establish properties of the inventory and advertising policy minimizing the expected discounted cost over a finite horizon in a dynamic nonstationary inventory model with random demand which is influenced by the level of goodwill. Under linearization of the cost associated with the maximum inventory and the advertising effect on demand, the model is shown to be equivalent to an inventory model with disposal. Many results of this paper are extended to cover convex ordering cost of inventory and time lag in delivery of stocks.

1. INTRODUCTION

The problem of optimal inventory management when the demand is partially controllable is the main focus of this paper. Few authors have attempted to analyze traditional inventory control models where the level of demand is a choice variable; generally, these attempts were concentrated on models where the demand is a function of the inventory level. One way to model the control one can exercise on the demand is to specify that the demand is an increasing function of advertising expenditures. Along those line, even fewer authors have attempted to optimally integrate inventory policy with advertising policy. In fact, the only reference we know of is Miercourt [9]. In this paper, we shall integrate them in the context of a discrete time dynamic inventory model involving a single commodity with random demand, which depends on the level of advertising. Under certain restrictions, the solution to the inventory problem with advertising is equivalent to Fukuda's [4] solution to the inventory model with disposal. Also Veinott [20] and Ignall and Veinott [5] have established the existence and the characterization of the optimal policy for a dynamic, nonstationary, multiproduct inventory, of which many of our results in Section 3 could be regarded as applications. Topkis' [17] and Veinott's [23] work on subadditive functions on sublattices and its application to inventory problems by Veinott [22] provides us with the proper tools and methods for our analysis. A knowledge of their theory, at least to the extent of the Appendix, is essential to understand this paper. An abundant literature (see Balcer [2]) covers the problem of optimizing advertisement expenditures, given a known demand-advertising relationship and no inventories. Most of it generalizes the model of Arrow and Nerlove [1], which introduces the concept of goodwill increasing with advertising and decaying exponentially over time.

The problem faced by the manager of a store with important inventory costs and an advertising budget is the theme of this paper. These problems are important because not only can jointly managed advertisement and inventory policies reduce costs, but such policies,

independently managed, may very well reduce profits, well below the level obtained in the absence of advertisement. Though the present model is developed on the theme of a store manager in charge of advertising and inventory, the basic equation could just as well be reinterpreted as trying to maintain the inventory at a positive level, like maintaining a work force or a number of working pieces of equipment at a prescribed level (let S be convex with a minimum at the prescribed level). In the example of the work force, the demand is simply the random number of individuals who will quit minus $g(b)$, the number prevented from doing so or rehired, as compared to the number of new workers, $y - x$, obtained outside of the organization.

In Section 2, we set up an inventory model with advertisement. In the next section, we present and interpret conditions for the existence of an optimal solution. In Section 4, the optimal solution is characterized and shown to be nondecreasing in the initial inventory and advertisement. In Section 5, results on the monotonicity of the optimal solution are extended. In Section 6, the problem is reduced to an inventory model with generalized disposal. By generalized disposal, we mean that purchasing and disposing of inventory could both take place concurrently at the beginning of every period. When advertisement and ordering costs are linear, this is equivalent to the usual disposal problem discussed by Fukuda [4].

In Section 7 many results of Section 4 are extended to a model where convex inventory ordering cost is allowed. In Section 8, under additional assumptions, the model with lag in delivery and promotion is shown to have the same properties as the model with no lag.

2. MODEL

In this paper, we will study a discrete time dynamic model of single commodity management, when the nonnegative demand for the commodity in each period is uncertain, but has a known distribution depending on the existing goodwill. At the beginning of each period, the manager knows the initial inventory x and goodwill $a \geqslant 0$, the present and future demand distributions, and the cost structure. He decides to instantaneously increase the initial inventory x to a level $y \geqslant x$ by ordering at a unit cost c, and the initial goodwill a to a level $b \geqslant a$ by advertising at a unit cost p. The random demand $g(b) + U$ depends on the goodwill and a nonnegative random number U. We assume that $g(0) = 0$ without loss of generality since a constant can always be added to U. During this period, the demand $g(b) + U$ occurs, so the terminal inventory becomes $y - g(b) - U$ and the terminal goodwill is unchanged. The initial inventory and goodwill in the next period are respectively $\eta[y - g(b) - U]$ and θb; $\theta \geqslant 0$ and $\eta(z) = \eta_+ z^+ - \eta_- z^-$, where $\eta_+ \geqslant \eta_- \geqslant 0$, $z^+ = \max(z, 0)$, and $z^- = \max(-z, 0)$. When the slope of η is between zero and one, $(1 - \eta_+)$ is interpreted as a depletion or loss of inventory and $(1 - \eta_-)$ as a loss of sales arising from the impatient consumers who depart before receiving their orders. Values of θ less than one correspond to depletion of consumer goodwill. This assumption corresponds to Zielske's [24] finding that the goodwill declines almost exponentially over time, and to Tull's [18] evidence on the carryover effect of advertisement. We introduce an additional concept, the book value of the terminal inventory, $y - g(b) - U$, and of the terminal goodwill, b, denoted $T[y - g(b) - U, b]$. The function $T(z,b)$ is equal to $-\lambda c'\eta(z) - \lambda p'\theta b$, where $\lambda \geqslant 0$ is the discount factor and the primes indicate the cost functions associated with the following period. If z, the present terminal stock, is positive, the next period initial stock is $\eta_+ z$, and the manager would pay the then discounted price $\lambda c'\eta_+ z$ to reach inventory level $\eta_+ z$, if no carryover of inventory were permissible from one period to the next. If z is negative, the next period initial stock is $\eta_- z$, and the manager would pay the then discounted price $\lambda c'\eta_- z^-$ to satisfy the backlogged demand completely. The next period initial

goodwill is θb, and the manager would pay the then discounted price $\lambda p'\theta b$ to reach that goodwill level, if no carryover of goodwill were possible from one period to the next. The book value of the last period terminal inventory and goodwill is called the terminal cost.

During each period, the manager incurs a capacity cost h on the starting inventory and a shortage-holding cost s on the terminal inventory. Both of these cost functions are convex, and the function s increases to infinity with its argument. Moreover, $s(z)$ is nondecreasing in z on the nonnegative real line. The function h is bounded below. There is a unit selling price r. The current sale price is paid when each consumer demand is incurred. This yields a gross revenue of $r[g(b) + U]$ to the manager. If the inventory is positive, the consumer receives the commodity without delay until either the demand is totally satisfied or the inventory is completely exhausted, whichever comes first. If consumers subsequently depart or increase their orders without being served, the manager refunds or pockets, respectively, the then current sale price. The sale is final only when consumers receive the commodity they have purchased. We assume that EU and $Es(y - g(b) - U)$ are finite, where E denotes the expectation, and that the demand function $g(b)$ is increasing and concave in the goodwill. This last assumption has been verified empirically by Shryer [12] and Stone [15] for mail ordering, Telser [16] for cigarettes, Palda [10] for drugs, Clement et al. [3] for milk, and Simon [13,14] for liquor.

Given a finite horizon N, let $\tilde{C}^n(x,a)$ be the minimum expected discounted cost in the periods $n, \ldots, N$, where x and a are, respectively, the initial inventory and goodwill in period n. The function $\tilde{C}^n$ can be calculated for each period by the dynamic programming recursion

$$\tilde{C}^n(x,a) = \min_{y \geq x, b \geq a} \{c(y - x) + p(b - a) + h(y) + Es[y - g(b) - U]$$

$$- r[g(b) + EU] + (1 - \eta_-)rE[y - g(b) - U]^- \tag{1}$$

$$+ \lambda E\, \tilde{C}^{n+1}[\eta(y - g(b) - U),\ \theta b]\}$$

for $n = 1, \ldots, N$, where $\tilde{C}^{N+1}(x,a) = -cx - pa$. Every symbol in Eq. (1) should be indexed by n; however, when no confusion is possible because of the context, the index n is suppressed throughout this paper. For simplicity, we will assume that the total demand is uncorrelated from one period to the next. However, the results to be proven in this paper can easily be generalized if the basic demands in a given period are correlated with the basic demands in a preceding period, independently of the goodwill.

It is convenient first to transform the problem to one with no ordering cost, a technique which has been used by Veinott [20]. On letting $C^n(x,a) = \tilde{C}^n(x,a) + cx + pa$, the recursion (1) becomes

$$C^n(x,a) = \min_{y \geq x, b \geq a} \{L(y,b) + \lambda EC^{n+1}[\eta(y - g(b) - U),\ \theta b]\}, \tag{2}$$

where

$$L(y,b) = cy + h(y) + (p - \lambda\theta p')b + Es\,[y - g(b) - U] - r[g(b) + EU] \tag{3}$$

$$+ [r - \eta_-(r - \lambda c')]\, E[y - g(b) - U]^- - \lambda c'\eta_+ E[y - g(b) - U]^+.$$

The term in braces in Eq. (2) is denoted $B^n(y,b)$. Under this transformation, $C^{N+1}(x,a) = 0$.

Upon regrouping the terms in Eq. (3), we obtain

$$L(y,b) = H(y) + P[g(b)] + S[y - g(b)], \tag{4}$$

where

$$H(z) = cz + h(z), \tag{5}$$

$$P(z) = (p - \lambda\theta p')g^{-1}(z) - rz, \tag{6}$$

and

$$S(z) = E[s(z - U) - \lambda c'\eta(z - U) + (1 - \eta_-)r(z - U)^-]. \tag{7}$$

A constant term $-rEU$ has been omitted as it clearly does not affect the choice of the optimal policy. The term $B^n(y,b)$ in Eq. (2) is not necessarily convex since it is the sum of the convex functions with the concave function $-\eta$. Hence we will transform Eq. (1) by replacing $g(b)$, $g(a)$, $C^n(x,a)$, $B^n(y,b)$, and $L_n(y,b)$ with β, α, $C^n(x,\alpha)$, $B^n(y,\beta)$, and $L_n(y,\beta)$, respectively. For simplicity, the same notation is used for C, L, and B whether the goodwill effect (α and β) or the advertising level (a and b) is used. The resulting recursion is

$$C^n(x,\alpha) = \min_{y \geqslant x, \beta \geqslant \alpha} \{H(y) + P(\beta) + S(y - \beta) + \lambda EC^{n+1}[\eta(y - \beta - U),\ g(\theta g^{-1}(\beta))]\}. \tag{8}$$

Because $g(b)$ is increasing in b and $g(b) = \beta$, the optimal policy $\bar{\beta}(x,\alpha)$ associated with recursion (8) can be expressed in terms of the optimal policy $\bar{b}(x,a)$.

In the new problem described by recursion (8), the variable β can be thought of as the additional guaranteed demand that the manager purchases at a price $P(\beta)$. This situation can arise when the manager discounts the merchandise to attract additional customers or to increase the quantity purchased by customers. In the remainder of this paper, the term goodwill will instead be called advertising or promotion. The term advertising is used when the presence of goodwill in one period influences the level of subsequent demands, i.e., $\theta > 0$, and the term promotion is used otherwise, i.e., $\theta = 0$. Thus the effect of advertising is persistent while that of promotion is ephemeral.

From here on, $(\bar{y}^n(x,\alpha),\ \bar{\beta}^n(x,\alpha))$ is the optimal policy in period n whose components are the optimal inventory and the optimal controlled demand and which minimizes the right-hand side of Eq. (8). The solution to the minimization of the total costs as given by Eq. (8) when only one of the two variables can be chosen and when that variable is unconstrained is denoted $\tilde{y}^n(\alpha)$ and $\tilde{\beta}^n(x)$, respectively. Also, the solution to the minimization of the total costs in period n as described by Eq. (8) when the two variables are unconstrained is $(y^{*n},\ \beta^{*n})$. This defines the base stock level in period n (as we shall see in Section 3) whose components are the base inventory level and the base controlled demand. The differences $\bar{y}^n(x,\alpha) - \bar{\beta}^n(x,\alpha)$ and $y^{*n} - \beta^{*n}$, are the optimal net inventory and the base net inventory in period n, rewritten $\bar{z}^n(x,\alpha)$ and z^{*n}, respectively. Finally, the solution to the minimization of the total present costs in period n, as given by the right-hand side of Eq. (8) with the last term omitted, is the myopic policy. In the preceding sentences, if we replace superscript n by subscript n, we have the myopic counterparts of the optimal solutions. For example, z_n^* is the myopic base net inventory in period n.

3. EXISTENCE OF OPTIMAL SOLUTIONS

Before proving the existence of an optimal solution, we will further specify the admissible functions g. When $\theta > 0$, we also assume that $\bar{g}(\beta) \equiv g(\theta g^{-1}(\beta))$ is convex in β. (The

reason for this will become clear shortly.) This means that each additional unit of demand purchased in a period generates more demand in the following period than the preceding units of demand. Also, since g is increasing, θ is positive and $g^{-1}(\beta)$ is positive when β is positive, it follows that $\bar{g}(\beta) < \beta$ for $\theta < 1$, $\bar{g}(\beta) = \beta$ for $\theta = 1$, and $\bar{g}(\beta) > \beta$ for $\theta > 1$.

Let Γ be the class of all concave increasing functions g such that $g(0) = 0$ and $\bar{g}(\beta)$ is convex in β. The class Γ' of all functions kb^{ν}, where $k > 0$ and $1 \geqslant \nu \geqslant 0$, is a subclass of Γ. Moreover, $\bar{g}(\beta)$ is linear in β for all elements of Γ'. If $\theta \leqslant 1$ and $k > 0$, $k \ln(b + 1)$ belongs to Γ.

To ensure that the minimand in Eq. (8) is convex, we assume that the following conditions hold hereafter:

CONDITION 1: $D^{+}s(0) + r - \lambda c'\eta_{+} \geqslant \eta_{-}(r - \lambda c') + D^{-}s(0)$,

CONDITION 2: $p \geqslant \lambda\theta p'$,

where $D^{+}s(0)$ ($D^{-}s(0)$) indicates the right-(left-) hand derivative of s at zero. The left-hand side of Condition 1 can be regarded as the marginal revenue generated by a guaranteed demand of size ϵ for a given inventory level ϵ, where ϵ is an arbitrarily small positive number. The marginal revenue consists of the marginal sales price minus the marginal book value of the inventory sold plus the marginal savings on holding cost when reducing the inventory level. The right-hand side term can be thought of as the marginal revenue generated by an arbitrarily small guaranteed demand, given a nonpositive inventory level. The marginal revenue consists of the marginal sales price on preserved sales minus the marginal book value of the inventory sold minus the marginal increase in shortage cost when reducing the inventory level. Therefore Condition 1 means that the marginal revenue at an arbitrarily small positive inventory level, generated by an equally small guaranteed demand, must be no smaller than the marginal revenue at any nonpositive inventory level, generated by an arbitrarily small guaranteed demand. Condition 2 asserts that the present price p of a unit of goodwill is greater than its book value.

Thus, both S and P are convex, since g is concave and increasing and therefore g^{-1} is convex and increasing. By the assumptions of Section 2, H is convex and P is continuous at zero. From this point on, we will deal with a problem with convex costs H, P, and S such that P is continuous at zero. We now show that

PROPOSITION 1: Under Conditions (1,2), $C^{n}(x,\alpha)$ is convex and nondecreasing in (x,α).

PROOF: Of course $C^{N+1} \equiv 0$ is convex. By the induction hypothesis, $C^{n+1}(x,\alpha)$ is convex in (x,α). Hence, because nondecreasing convex functions of convex functions are convex, $\eta(y - \beta - U)$ and $EC^{n+1}[\eta(y - \beta - U), \bar{g}(\beta)]$ are convex in (y,β). Thus $B^{n}(y,\beta)$ is convex in (y,β). The minimum of a convex function over a convex set being convex, $C^{n}(x,\alpha)$ is convex in (x,α). Because the sets over which the minimization takes place are decreasing in (x,α), the function $C^{n}(x,\alpha)$ is nondecreasing in (x,α). This concludes the proof.

Since $B^{n}(y,\beta)$ is convex and convex functions are continuous, $B^{n}(y,\beta)$ is continuous. Since $B^{n}(y,\beta)$ increases to infinity with $|y| + \beta$ under conditions to be described, the space over which the minimization takes place can be limited to a compact set. Also, the intersection

of a closed set $\{y \geqslant x,\ \beta \geqslant \alpha\}$ of R^2 with a compact set is compact. Since the minimizing set of a continuous function over a compact set is compact, the minimizing set $M(x,\alpha)$ of $B^n(y,\beta)$ over $\{y \geqslant x,\ \beta \geqslant \alpha\}$ has a lexicographically least element $(\bar{y}(x,\alpha), \bar{\beta}(x,\alpha))$, called the optimal policy.

We proceed to show that $B^n(y,\beta)$ tends to infinity with $|y| + \beta$ under the following additional conditions, which are assumed to hold hereafter in this paper:

CONDITION 3: $D^-H(\infty) + D^-S(\infty) > 0$,

CONDITION 4: $D^-H(\infty) + D^-P(\infty) > 0$,

CONDITION 5: $D^-P(\infty) - D^+S(-\infty) > 0$,

CONDITION 6: $D^+H(-\infty) + D^+S(-\infty) < 0$.

The interpretation of Conditions (3-6) is very natural. Condition (3) states that with fixed added demand, it is not profitable to increase the starting inventory without limit. Condition (4) states that it is not profitable to increase together indefinitely the added demand and the inventory. Condition (5) states that when no inventory is ordered, it does not pay to increase the added demand indefinitely. Finally, Condition (6) states that with fixed added demand, it does not pay to let the inventory reach a very small level.

PROPOSITION 2: Under Conditions (1,2), the function $L_n(y,\beta)$ tends to infinity with $|y| + \beta (\beta \geqslant 0)$ if and only if Conditions (3-6) hold.

PROOF: The only if part follows by observing that since $L_n(\cdot,\cdot)$ is convex, Conditions (3-6) are respectively equivalent to the assertion that $L_n(y,\beta) \to \infty$ along the respective half lines $y \geqslant 0, \beta = 0$; $y = \beta \geqslant 0$; $y = 0, \beta \geqslant 0$; and $y \leqslant 0, \beta = 0$.

For the converse, recall first that the convex function $L_n(y,\beta) \to \infty$ as $|y| + \beta \to \infty$ if that is so along every half line emanating from the origin. Let $\beta = \delta y$, so $L_n(y,\beta)$ associated with Eq. (5) becomes

$$\phi(y) \equiv L_n(y,\delta y) = H(y) + P(\delta y) + S((1-\delta)y)\,.$$

CASE 1: If $y \geqslant 0$ and $1 \geqslant \delta \geqslant 0$, then

$$\begin{aligned}\lim_{y\to\infty} D^-\phi(y) &= \lim_{y\to\infty} \{D^-H(y) + \delta D^-P(\delta y) + (1-\delta)D^-S((1-\delta)y)\} \\ &= \lim_{y\to\infty} \{\delta[D^-H(y) + D^-P(\delta y)] + (1-\delta)[D^-H(y) + D^-S((1-\delta)y)]\} \\ &= \delta[D^-H(\infty) + D^-P(\infty)] + (1-\delta)[D^-H(\infty) + D^-S(\infty)] > 0\end{aligned}$$

since Conditions (3,4) hold.

CASE 2: If $y \geqslant 0$ and $\delta \geqslant 1$, then

$$\begin{aligned}\lim_{y\to\infty} D^-\phi(y) &= \lim_{y\to\infty} \{D^-H(y) + \delta D^-P(\delta y) + (1-\delta)D^+S((1-\delta)y)\} \\ &= \lim_{y\to\infty} \{[D^-H(y) + D^-P(\delta y)] + (\delta-1)[D^-P(\delta y) - D^+S((1-\delta)y)]\} \\ &= [D^-H(\infty) + D^-P(\infty)] + (\delta-1)[D^-P(\infty) - D^+S(-\infty)] > 0\end{aligned}$$

since Conditions (4,5) hold.

CASE 3: Let $y \leqslant 0$ and $\delta \leqslant 0$, then

$$\lim_{y\to-\infty} D^+\phi(y) = \lim_{y\to-\infty} \{D^+H(y) + \delta D^-P(\delta y) + (1-\delta)D^+S((1-\delta)y)\}$$

$$= \lim_{y\to-\infty} \{[D^+H(y) + D^+S((1-\delta)y)] + \delta[D^-P(\delta y) - D^+S((1-\delta)y)]\}$$

$$= [D^+H(-\infty) + D^+S(-\infty)] + \delta[D^-P(\infty) - D^+S(-\infty)] < 0$$

since Conditions (5,6) hold. This completes the proof of Proposition 2.

PROPOSITION 3: If $L_i(y,b)$ tends to infinity with $|y| + b$ for $i = n, \ldots, N$, then $B^n(y,b)$ tends to infinity with $|y| + b$.

PROOF: This is obviously true for B^N since $B^N = L_N$. By the induction hypothesis, B^{n+1} has the property. Hence, by convexity B^{n+1} is bounded below on $R \times R_+$. Thus, C^{n+1} is bounded below. From this, we conclude that $B^n(y,b)$ tends to infinity with $|y| + b$. This proves the assertion.

Proposition 3 completes the proof of the existence of the optimal policy.

In the rest of this section, we will establish sufficient conditions to warrant the existence of nonnegative base inventory levels. When the base inventory levels are nonnegative, the restriction $\eta_- \leqslant \eta_+$ can be waived since the optimal policy $(\bar{y}(x,\alpha), \bar{\beta}(x,\alpha))$ is greater (see Theorem 1 in Section 4) than or equal to $(y^*, \beta^*) \geqslant (0,0)$ for all (x,α), and since $D_1^-C^{n+1}[\eta(w), \bar{g}(v)] = 0$ for all $w \leqslant 0$ and $v \geqslant 0$, where D_i^-C denotes the left-hand derivation of the function C with respect to its i th argument. Obviously, Conditions (1-6) must still hold.

We assert that if the myopic base inventory level for each period is nonnegative, then the base inventory level is also nonnegative in each period. We show this by induction. Since C^{N+1} is zero, the myopic base inventory level in period N is equal to the period N base inventory level. By induction, the result holds for period $n+1$. Therefore, $D_1^-C^{n+1}(x,\alpha) = 0$ on $R_- \times R_+$. Since $\eta(x) \leqslant 0$ for $x \leqslant 0$, θ is nonnegative, and the demand is nonnegative, $ED_1^-C^{n+1}[\eta(y-\beta-U), \bar{g}(\beta)] = 0$ for all $(y,\beta) \in R_- \times R_+$. This, with the fact that the myopic base inventory level is nonnegative in period n, implies that the period n base inventory level is nonnegative, completing the proof of the assertion.

An expedient method to guarantee nonnegative base inventory levels is to impose the restriction $y \geqslant 0$. Another method is to impose an additional condition on the cost structure such that $\tilde{y}(0) \geqslant 0$, which is sufficient to ensure that $y^* \geqslant 0$, since $0 \leqslant \tilde{y}(0) \leqslant \tilde{y}(\beta^*) = y^*$ (see Theorem 1) where $\beta^* \geqslant 0$. Because of the preceding paragraph, it will suffice to establish the result for myopic base inventory levels. By differentiating Eq. (4) with respect to y with $b = \beta = 0$, we obtain the desired condition

CONDITION 7: $D^-H(0) + D^-S(0) < 0$,

which ensures nonnegative base solutions. For inventory problems with perishable goods, i.e., $\eta_+ = 0$, the nonnegative base inventory guarantees that the optimal policy is independent of the level of backlogged demand, $\eta_- x^-$.

4. CHARACTERIZATION OF THE OPTIMAL POLICY

In this section, we will first establish that $C^n(x,\alpha)$ is subadditive in (x,α), then characterize the minimizing set $M^n(x,\alpha)$ and the form of the optimal solution. We remind the reader that the starting stock is (x,α).

PROPOSITION 4: If Conditions (1-6) hold, $C^n(x,\alpha)$ is subadditive in (x,α) and $B^n(y,\beta)$ is subadditive in (y,β).

PROOF: Clearly, $C^{N+1} \equiv 0$ is subadditive in (x,α). By the induction hypothesis, $C^{n+1}(x,\alpha)$ is subadditive in (x,α). Therefore, since η is convex nondecreasing and $\bar{g}$ is nondecreasing, $EC^{n+1}[\eta(y-\beta-U), \bar{g}(\beta)]$ is subadditive in (y,β) by Example A7. (This and the following examples can be found in the Appendix). Because S is convex, $S(y-\beta)$ is subadditive in (y,β) by Example A6. Thus, by Example A8, $B^n(y,\beta)$ is subadditive in (y,β). Since $B^n(y,\beta)$ is independent of (x,α), $B^n(y,\beta)$ is subadditive in (x,α,y,β). Also, $\{y \geqslant x,\ \beta \geqslant \alpha\}$ is a sublattice of R^4 by Example A1. Since the minimum of a subadditive function over a sublattice is subadditive by Theorem A1, (also in the Appendix) $C^n(x,\alpha)$ is subadditive in (x,α), completing the proof of the proposition.

We now turn our attention to the set $M^n(x,\alpha)$.

PROPOSITION 5: The minimizing set $M^n(x,\alpha)$ is a nonempty compact sublattice.

PROOF: By Proposition 1, we know that $B^n(y,\beta)$ is continuous. By Proposition 4, we also know that $B^n(y,\beta)$ is subadditive in (y,β). Also $B^n(y,\beta)$ tends to infinity with $|y|+\beta$. Therefore, by Theorem A2 (see Appendix) the minimizing set $M^n(x,\alpha)$ is a nonempty compact sublattice. This completes the proof of the assertion.

The lexicographically least element of $M^n(x,\alpha)$ is then the least element of $M^n(x,\alpha)$ and is called the optimal policy, written $(\bar{y}(x,\alpha), \bar{\beta}(x,\alpha))$. Also, the least element of $M^n(-\infty,0)$ is denoted (y^*,β^*) and called the base stock level. Before further characterizing the optimal policy, we recall that $\tilde{y}(\alpha)$ is the least y minimizing $B^n(y,\beta)$ subject to $\beta = \alpha$ and $\tilde{\beta}(x)$ is the least β minimizing $B^n(y,\beta)$ subject to $y = x$ and $\beta \geqslant 0$.

THEOREM 1: Under Conditions (1-6), $\bar{y}(x,\alpha) = x \vee \tilde{y}(\beta^* \vee \alpha)$ and $\bar{\beta}(x,\alpha) = \alpha \vee \tilde{\beta}(y^* \vee x)$ with $\tilde{y}$ and $\tilde{\beta}$ being nondecreasing on their respective domains.

PROOF: Proposition 5 and Theorem A2 imply that the optimal policy exists and is nondecreasing in (x,α) and that $\tilde{\beta}$ and $\tilde{y}$ are nondecreasing. Also $(\tilde{y}(\beta^*), \tilde{\beta}(y^*)) = (y^*,\beta^*)$. Now for $(x,\alpha) \leqslant (y^*,\beta^*)$, $(\bar{y}(x,\alpha), \beta(x,\alpha)) = (y^*,\beta^*)$. For $x \geqslant y^*$ and $\alpha \leqslant \tilde{\beta}(x)$, we have by the convexity of B^n that $\bar{y}(x,\alpha) = x$ and so $\bar{\beta}(x,\alpha) = \tilde{\beta}(x)$. Similarly, for $\alpha \geqslant \beta^*$ and $x \leqslant \tilde{y}(\alpha)$, $\bar{y}(x,\alpha) = \tilde{y}(\alpha)$ and $\bar{\beta}(x,\alpha) = \alpha$. Combining these two facts we see that if $(y^*,\beta^*) \leqslant (x,\alpha) \leqslant (\tilde{y}(\alpha), \tilde{\beta}(x))$, then $(x,\alpha) = (\bar{y}(x,\alpha), \bar{\beta}(x,\alpha))$. Now if $(y^*,\beta^*) \leqslant (x,\alpha)$ and if $(\tilde{y}(\alpha), \tilde{\beta}(x)) \leqslant (x,\alpha)$, then $(\bar{y}(x,\alpha), \bar{\beta}(x,\alpha)) = (x,\alpha)$. For if not, there are three possibilities, viz., (i) $\bar{y}(x,\alpha) > x$ and $\bar{\beta}(x,\alpha) = \alpha$, (ii) $\bar{y}(x,\alpha) = x$ and $\bar{\beta}(x,\alpha) > \alpha$, or (iii) $\bar{y}(x,\alpha) > x$ and $\bar{\beta}(x,\alpha) > \alpha$. If (i) occurs, then $\tilde{y}(\alpha) = \bar{y}(x,\alpha) > x \geqslant \tilde{y}(\alpha)$ which is impossible. If (ii) holds, then $\tilde{\beta}(x) = \tilde{\beta}(x,\alpha) > \alpha \geqslant \tilde{\beta}(x)$ which is also a contradiction. Finally, if (iii) holds, then since B^n is convex, B^n is nonincreasing along the line from $(\bar{y},\bar{\beta})$ to (y^*,β^*), implying that either (i) or (ii) holds, contradicting (iii). This completes the proof.

To sum up, we have shown that the optimal policy is as depicted in Figure 1. Also, we have proved $\tilde{y}(\beta^*) = y^*$ and $\beta^* = \tilde{\beta}(y^*)$.

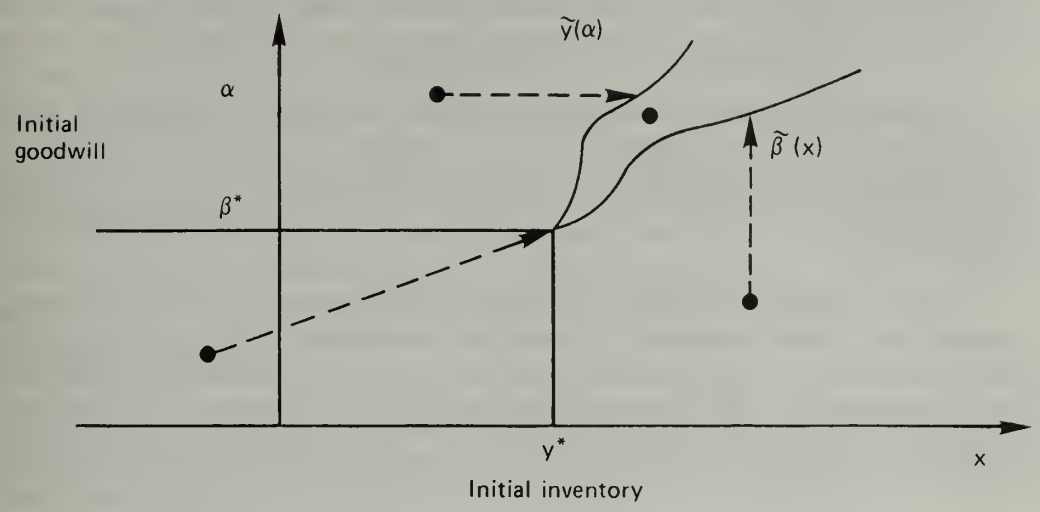

FIGURE 1. — Depiction of optimal policy.

For this paragraph, we further assume that the costs and the demands are stationary and that the depletion factors of inventory, η_+, and of goodwill, θ, are no greater than one. When the initial inventory x and goodwill α are smaller than the base stock level of the first period, the myopic policy is optimal in every period [20]. It is optimal to order $(y^* - x, \beta^* - \alpha)$ at the beginning of the period. By further restricting λ to be smaller than one, the model easily carries over to the infinite horizon. We could obtain myopic optimal policy for this problem under a nonstationary structure by applying the conditions defined in Veinott [20] and in Ignall and Veinott [5].

5. MONOTONICITY OF THE OPTIMAL POLICY

In this section, we shall take advantage of the particular structure of the optimal policy to exhibit its monotonicity in the initial parameters. Namely, we will prove:

THEOREM 2: For all n, $\bar{y}^n(x,\alpha)$ and $\bar{\beta}^n(x,\alpha)$ are nondecreasing in (x,α). Also $\bar{z}^n(x,\alpha)$, $x - \bar{y}^n(x,\alpha)$, $x - \bar{\beta}^n(x,\alpha)$, and $x - \bar{z}^n(x,\alpha)$ are nondecreasing in x. Moreover, $\alpha - \bar{\beta}^n(x,\alpha)$ and $\alpha + \bar{z}^n(x,\alpha)$ are nondecreasing in α. Finally, if H is linear, $\alpha - \bar{y}^n(x,\alpha)$ is unimodal in α, and $\bar{z}^n(x,\alpha)$ is nondecreasing in α.

PROOF: By Theorem 1, $\bar{y}^n(x,\alpha)$ and $\bar{\beta}^n(x,\alpha)$ are nondecreasing in (x,α). Let $y = x$ and $\beta' = x - \beta$, and consider

$$\min_{x-\beta' \geqslant 0} \{P(x-\beta') + S(\beta') + \lambda EC^{n+1}[\eta(\beta' - U),\ \bar{g}(x-\beta')]\}.$$

Since P is convex, $P(x-\beta')$ is subadditive in (x,β') by Example A6, and since C^{n+1} is subadditive on $R \times R_+$, η and $\bar{g}$ are nondecreasing. $\lambda EC^{n+1}[\eta(\beta' - U), \bar{g}(x-\beta')]$ is subadditive in (x,β') by Example A7. By Theorem A2, the greatest element $\beta' = x - \tilde{\beta}^n(x)$, minimizing the erm inside the braces subject to $\beta' \leqslant x$, is nondecreasing in x. Thus, $x - \bar{y}^n(x,\alpha) = x - x \vee \tilde{y}^n(\beta^{*n} \vee \alpha)$ and $x - \bar{\beta}^n(x,\alpha) = x - \alpha \vee \tilde{\beta}^n(y^{*n} \vee x)$ are nondecreasing in x. Also, $\bar{z}^n(x,\alpha) = x \vee \tilde{y}^n(\beta^{*n} \vee \alpha) - \alpha \vee \tilde{\beta}^n(y^{n*} \vee x)$ is nondecreasing in x. Moreover, $x - \bar{z}^n(x,\alpha) = x - \bar{y}^n(x,\alpha) + \bar{\beta}^n(x,\alpha)$ is nondecreasing in x.

Now, $\alpha - \bar{\beta}^n(x,\alpha) = \alpha - \alpha \vee \tilde{\beta}^n(y^{*n} \vee x)$ is nondecreasing in α. Also, $\alpha + \bar{z}^n(x,\alpha) = \alpha - \bar{\beta}^n(x,\alpha) + \bar{y}^n(x,\alpha)$ is nondecreasing in α. Let $\beta = \alpha$ and $-y' = \alpha - y$, and consider

$$\min_{y'} \{H(\alpha + y') + S(y') + \lambda EC^{n+1}[\eta(y' - U), \bar{g}(\alpha)]\}.$$

If H is linear, the term inside the braces is subadditive in (α, y') by Examples A5 and A8 and the fact that $C^{n+1}(x,\alpha)$ is subadditive in (x,α). Therefore, the least $y' = \tilde{y}^n(\alpha) - \alpha$ achieving the minimum is nondecreasing in α. Thus, $\alpha - \bar{y}^n(x,\alpha) = \alpha - x \vee \tilde{y}^n(\beta^{*n} \vee \alpha)$ is unimodal in α. Since $\bar{z}^n(x,\alpha) = x \vee \tilde{y}^n(\beta^{*n} \vee \alpha) - \alpha \vee \tilde{\beta}^n(y^{*n} \vee x)$, $\bar{z}^n(x,\alpha)$ is nondecreasing in α, completing the proof.

When the goods are assumed to be totally perishable, i.e., $\eta_+ = 0$, we can transform the problem into a convex program and exhibit the monotonicity of the optimal policy in the initial guaranteed demand α. By imposing Condition (7) or $y \geqslant 0$, we ensure that the base inventory level is nonnegative. Since the goods are perishable, $x \leqslant 0$ always. Also, since y^{*n} is always nonnegative, $C^n(x,\alpha) = C^n(0,\alpha)$ for all $x \leqslant 0$. Thus, with $C^n(\alpha) \equiv C^n(0,\alpha)$ and $\eta_+ = 0$, recursion (8) becomes

$$C^n(\alpha) = \min_{y \geqslant 0, \beta \geqslant \alpha} \{H(y) + P(\beta) + S(y - \beta) + \lambda C^{n+1}[\bar{g}(\beta)]\}. \tag{9}$$

Since the state variable in Eq. (9) is not random, it follows that the advertising of perishable goods model is a deterministic convex program, viz.,

$$C^1(\alpha) = \min \sum_{n=1}^{N} [H_n(y_n) + P_n(\beta_n) + S_n(y_n - \beta_n)] \tag{10}$$

such that $y_n \geqslant 0$ for all n, $\beta_n \geqslant \bar{g}(\beta_{n-1})$ for all $n > 1$, and $\beta_1 \geqslant \alpha$. Thus it is not necessary to find the optimal policy for all values of the initial controlled demand in every period. Since the right-hand side of Eq. (10) is subadditive in y_n and β_n, and the set of feasible solutions is a nonempty closed sublattice of R^{2N} that is bounded below, there is a least vector $(\bar{y}_1^1(\alpha), \ldots, \bar{y}_N^1(\alpha), \bar{\beta}_1^1(\alpha), \ldots, \bar{\beta}_N^1(\alpha))$ minimizing the right-side of Eq. (10) over the feasible region. These vectors can be found by nonlinear programming algorithms. Let $\bar{z}_n^1(\alpha) \equiv \bar{y}_n^1(\alpha) - \bar{\beta}_n^1(\alpha)$.

THEOREM 3: When $\eta_+ = 0$, for each n, the functions $\bar{y}_n^1(\alpha)$, $\bar{\beta}_n^1(\alpha)$, $-\bar{z}_n^1(\alpha)$, $\alpha - \bar{y}^n(\alpha)$, $\alpha - \bar{\beta}^n(\alpha)$, and $\alpha + \bar{z}^n(\alpha)$ are nondecreasing in α.

PROOF: By Theorem A2, $(\bar{y}_n^1(\alpha), \bar{\beta}_n^1(\alpha))$ is nondecreasing in $\alpha \geqslant 0$ for all n. Replacing $y_n - \beta_n$ by $-z_n'$, the right-hand side of Eq. (10) becomes

$$\min \sum_{n=1}^{N} [H_n(\beta_n - z_n') + P_n(\beta_n) + S_n(-z_n')] \tag{11}$$

subject to $\beta_n \geqslant z_n'$ and $\beta_n \geqslant \bar{g}(\beta_{n-1})$. Since H_n is convex, by the same argument as above, the $\{N+1, \ldots, 2N\}$-lexicographically least element $(\bar{\beta}_1^1(\alpha), \ldots, \bar{\beta}_N^1(\alpha), -\bar{z}_1^1(\alpha), \ldots, -\bar{z}_N^1(\alpha))$ minimizing Eq. (11) subject to the constraints indicated above is nondecreasing in α, so $\bar{z}_n^1(\alpha)$ is nonincreasing in α for all n.

If instead we let $\beta' = \alpha - \beta$ and $y' = \alpha - y$, Eq. (9) becomes

$$C^n(\alpha) = \min_{y' \leqslant \alpha, \beta' \leqslant 0} \{H(\alpha - y') + P(\alpha - \beta') + S(\beta' - y') + \lambda C^{n+1}[\bar{g}(\alpha - \beta')]\}. \tag{12}$$

Since P, H, S, and $C^{n+1}(\bar{g}(\cdot))$ are convex, the term inside the braces in Eq. (12) is subadditive in (α, y', β') by Example A6. The space over which the minimization takes place is a sublattice by Example A1. By Theorem A2, the greatest element $(\alpha - \bar{y}^n(\alpha), \alpha - \bar{\beta}^n(\alpha))$, minimizing the right hand side of Eq. (12), is nondecreasing in α. Moreover, $\alpha + \bar{z}^n(\alpha) = \alpha - \bar{\beta}^n(\alpha) + \bar{y}^n(\alpha)$ is also nondecreasing in α by the above, completing the proof.

The results of Theorem 3 are stronger than those of Theorem 2, as not only the current optimal policy is increasing in the initial parameter but also all subsequent optimal policies. As the guaranteed demand due to past advertisement increases, the additional impact of optimal current expenditures on advertisement must not increase. Similarly, the optimal inventory level must not increase as fast as guaranteed demand due to past advertisement.

6. GENERALIZED DISPOSAL

In the remaining sections, θ is assumed to be 0. The assumption $\theta = 0$ means that the effect of advertising is ephemeral and therefore the initial goodwill in each period is $\alpha = 0$. The problem as described by Eq. (8) is thus reduced to

$$C^n(x) = \min_{y \geqslant x, \beta \geqslant 0} \{H(y) + P(\beta) + S(y - \beta) + \lambda E C^{n+1}[\eta(y - \beta - U)]\}, \tag{13}$$

where $C^{n+1} \equiv 0$. Heretofore, we have supposed the cost $p(b)$ of purchasing b units of promotion in a period is linear in b, and that g and $\bar{g}$ are concave and convex, respectively, and nonnegative and increasing. From now on, we drop these assumptions and assume only that g is nonnegative and increasing, and that $p(g^{-1}(\beta))$ is convex in $\beta \geqslant 0$. Also, let $P(\beta) = p(g^{-1}(\beta)) - r\beta$. Thus, each additional unit of added demand in a period is at least as expensive as its predecessor. Under Conditions (3-6) with $D^- p(g^{-1}(\infty))$ replacing $(p - \lambda\theta p')D^- g^{-1}(\infty)$, an optimal solution will exist when the problem is expressed in the original costs.

We will show that under the linearity of P and H, the problem at hand is equivalent to the usual inventory model with disposal as discussed by Fukuda [4]. If P is linear the result of Theorem 1 can be sharpened in the following way. Let $S'(z) = S(z) - Pz$ and $z = y - \beta$ then Eq. (13) becomes

$$C^n(x) = \min_{x \geqslant z} \{H(x) + Px + S'(z) + \lambda E C^{n+1}[\eta(z - U)]\}, \tag{14}$$

so $\tilde{\beta}^n(x) = (x - z^{**n})^+$, where z^{**n} is the greatest z minimizing the sum of the last two terms in braces in Eq. (14) over R. This result is summarized below.

LEMMA 1: If P is linear on R_+, then $\tilde{\beta}^n(x) = (x - z^{**n})^+$, $\bar{\beta}^n(x) = (x - z^{**n})^+ \vee (y^{*n} - z^{**n}) = (x \vee y^{*n} - z^{**n})^+$, and $\bar{z}^n(x) = z^{**n} \wedge (x \vee y^{*n})$.

REMARK: If $y^{*n} \geqslant z^{**n}$, then $\bar{z}^n(x) = z^{**n} = z^{*n}$ for all x and if $y^{*n} < z^{**n}$, then

$$\bar{z}^n(x) = \begin{cases} y^{*n} = z^{*n} & \text{for } x \leqslant y^{*n} \\ x & \text{for } y^{*n} \leqslant x \leqslant z^{**n} \\ z^{**n} & \text{for } z^{**n} \leqslant x \end{cases}. \tag{15}$$

An important question here is: when is $y^{*n} \leqslant z^{**n}$? We already know that $y^{*n} \leqslant z^{**n}$ implies $\beta^{*n} = 0$. Therefore y^{*n} is the least y minimizing

$$\{H(y) + S(y) + \lambda E C^{n+1}[\eta(y - U)]\}. \tag{16}$$

By comparing Eqs. (14) and (16), we obtain that $D^{-}H(y^{*n}) + P \geqslant 0$ implies $y^{*n} \leqslant z^{**n}$. Similarly, a necessary condition for $y^{*n} \leqslant z^{**n}$ is $D^{+}H(y^{*n}) + P \geqslant 0$.

In the rest of this subsection we assume also that H is linear, so Eq. (13) becomes

$$C^n(x) = \min_{y \geqslant x, \beta \geqslant 0} \{Hy + P\beta + S(y - \beta) + \lambda EC^{n+1}[\eta(y - \beta - U)]\}.$$

On letting $z = y - \beta$, we have

$$C^n(x) = \min_{y \geqslant z \vee x} \{(H + P)y - Pz + S(z) + \lambda EC^{n+1}[\eta(z - U)]\}.$$

Since $H + P > 0$ by Condition (4), $y = z \vee x$ in the above recursion which then reduces to

$$C^n(x) = \min_z \{(H + P)(z \vee x) - Pz + S(z) + \lambda EC^{n+1}[\eta(z - U)]\}$$

$$= \min_z \{Hx + H(z - x)^+ + P(z - x)^- + S(z) + \lambda EC^{n+1}[\eta(z - U)]\}.$$

Let $\tilde{C}^n(x) = C^n(x) - Hx$ and $\tilde{S}(z) = S(z) + \lambda H'E\eta(z - U)$, so the new recursion is

$$\tilde{C}^n(x) = \min_z \{H(z - x)^+ + P(z - x)^- + \tilde{S}(z) + \lambda E\tilde{C}^{n+1}[\eta(z - U)]\}.$$

This is the usual formulation of the recursion for the inventory problem with disposal, where y^{*n} is the base ordering level and z^{**n} the base disposal level.

7. CONVEX PRODUCT ORDERING COST

In this section, we generalize our model from linear to convex product ordering cost. Suppose now that there is a nondecreasing ordering cost function $\tilde{c}$, where $\tilde{c}(z) = 0$ for $z \leqslant 0$. If we define c to be equal to $D^{+}\tilde{c}(0)$, then the right-hand derivative of the function $c(z) = \tilde{c}(z) - cz$ is equal to zero at zero. Under this transformation Eq. (13) becomes

$$C^n(x) = \min_{y \geqslant x, \beta \geqslant 0} \{c(y - x) + H(y) + P(\beta) + S(y - \beta) + \lambda EC^{n+1}[\eta(y - \beta - U)]\}, \tag{17}$$

where H, P, and S are defined by Eqs. (5), (6), and (7), and $C^{n+1}(x) \equiv 0$. Under Conditions (3-6), $L_n(y, \beta)$ tends to infinity with $|y| + \beta$ by Proposition 2. Since $c(y - x)$ is nonnegative for all y, $L_n(y,\beta,x) = L_n(y,\beta) + c(y - x)$ also tends to infinity with $|y| + \beta$ for every x. By Proposition 3, $B^n(y,\beta,x)$ does similarly upon noticing that $C^{n+1}(x)$ is bounded below. Therefore, an optimal solution exists for every x. As $L_n(y,\beta,x)$ is a convex function of (y,β,x), $C^n(x)$ is convex in x, provided η is linear.

THEOREM 4: If η is linear, $\bar{y}^n(x)$, $\bar{z}^n(x)$, $\bar{\beta}^n(x)$, $x - \bar{y}^n(x)$, $x - \bar{z}^n(x)$, and $x - \bar{\beta}^n(x)$ are each nondecreasing in x.

PROOF: Since the additional term $c(y - x)$ in Eq. (17) relative to Eq. (13) is subadditive in (y,x), by Proposition 5 $B^n(x,y,\beta)$ is subadditive in (x,y,β), and by Theorem A2 $\bar{y}^n(x)$ and $\bar{\beta}^n(x)$ are nondecreasing in x. Let $z = y - \beta$ in Eq. (17) which becomes

$$C^n(x) = \min_{y \geqslant x, y \geqslant z} \{c(y - x) + H(y) + P(y - z) + S(z) + \lambda EC^{n+1}[\eta(z - U)]\}. \tag{18}$$

Since P and c are convex, the term inside the braces is subadditive in (x,y,z) by Examples A6 and A8. Therefore, by Theorem A2 the 2-lexicographically least element, $(\bar{y}^n(x), \bar{z}^n(x))$ minimizing the right-hand side of Eq. (18) is nondecreasing in x.

Upon letting $y' = x - y$ and $\beta' = x - \beta$, Eq. (17) becomes

$$C^n(x) = \min_{0 \geqslant y', x \geqslant \beta'} \{c(-y') + H(x - y') + P(x - \beta') + S(\beta' - y') + \lambda E C^{n+1}[\eta(\beta' - y' - U)]\}. \tag{19}$$

Since H, P, S and C^{n+1} are convex and η is linear, the term inside the braces in Eq. (19) is subadditive in (x,y',β') by Examples A6 and A8. Therefore, by Theorem A2 the $\{1,2\}$-lexicographically least element $(x - \bar{y}^n(x), x - \bar{\beta}^n(x))$ minimizing the right-hand side of Eq. (19) is nondecreasing in x. Moreover, $x - \bar{z}^n(x) = (x - \bar{y}^n(x)) + \bar{\beta}^n(x)$ is nondecreasing in x by the above, completing the proof.

8. COMMON TIME LAG IN DELIVERY AND PROMOTION

In this section, we generalize our models by allowing a common interval of time to occur between the moments of product ordering and of promotion purchase, respectively, and the moments of the delivery of the product and of the promotion effect. This interval of time, ν, where ν is a nonnegative real integer, is called a time lag.

Let x be the inventory on hand plus on order minus the controlled demands of periods n to $n + \nu - 1$, at the beginning of period n. Also, β is the controlled demand that will take effect in period $n + \nu$. In addition, all costs are incurred at the beginning of period $n + \nu$. Under the assumption $\eta(x) = x$, recursion (13) becomes

$$C^n(x) = \min_{y \geqslant x, \beta \geqslant 0} \{H(y) + P(\beta) + S(y - \beta) + \lambda E C^{n+1}(y - \beta - U)\}, \tag{20}$$

where $n \leqslant N - \nu$, H, and P are defined by Eqs. (5) and (6), and

$$S(z) = Es(z - V) - \lambda c'z,$$

with V a random variable whose distribution is the convolution of $F_n, \ldots, F_{n+\nu}$.

Since the structure of Eq. (20) is identical to that of Eq. (13), all results discussed in Sections 2-6 apply directly.

9. CONCLUSION

This paper has shown the existence of an optimal policy to an inventory problem where the demand is influenced through advertising under the control of the same management unit. The present approach links together two generally separate functions of firm management, though there are examples of such a link in the real world, such as an ad campaign for an end of season sale, or an increase in inventory by merchants prior to a major ad blitz by a producer or by themselves. When the optimal policy for inventory level and advertising level is characterized, this link is rendered more explicit as each component of the optimal policy is increasing in the other and also in the initial levels of inventory and advertising.

Numerically, this two-variable dynamic program can be treated like two one-variable dynamic programs since the optimal policy $(x \vee \tilde{y}(\beta^* \vee \alpha), \alpha \vee \tilde{\beta}(x \vee y^*))$ is a known function of two single variable functions y and β over the range (y^*, ∞) and (β^*, ∞), respectively. Since, in general, these functions cannot be obtained in closed form, they are evaluated using one-variable search procedures. Thus, the number of numerical computations necessary to solve this problem are of the same order of magnitude as that of a one-variable dynamic program.

ACKNOWLEDGMENTS

The author thanks Arthur F. Veinott, Jr., who, as a teacher and thesis director, has devoted countless hours helping to refine many ideas. The author is particularly grateful for the privilege of using his work on lattice programming. He also gratefully acknowledges financial support from the Conseil des Arts du Canada. the Conseil National de Recherches du Canada, and the Office of Naval Research.

BIBLIOGRAPHY

[1] Arrow, K.J. and M. Nerlove, "Optimal Advertising Policy under Dynamic Conditions," Economica *29*, 129-142, (1962).

[2] Balcer, Y., "Optimal Advertising and Inventory Policy With Random Demands," unpublished Ph.D. Thesis, Operations Research, Stanford University (1974).

[3] Clement, W.E., P.L. Henderson and C.P. Eley, "The Effects of Different Levels of Promotional Expenditure on Sales of Fluid Milk," USDA, Economic Research Service, Washington, D.C. (1965).

[4] Fukuda, Y., "Optimal Disposal Policies," Naval Research Logistics Quarterly, *8*, 221-227 (1961).

[5] Ignall, E. and A.F. Veinott, Jr., "Optimality of Myopic Inventory Policies for Several Substitute Products," Management Science *15*, 284-304 (1969).

[6] Karlin, S., "Dynamic Inventory Policy With Varying Stochastic Demand," Management Science *6*, 231-258 (1960).

[7] Kuehn, A., "A Model for Budgeting Advertising," in Bass, F.M., *Mathematical Models and Methods in Marketing*, Homewood, Irwin (1961).

[8] Kuehn, A., "How Advertising Performance Depends on Other Marketing Factors," Journal of Advertising Research *2*, 2-10 (1962).

[9] Miercort, F.A., "Some Effects of Advertising and Prices on Optimal Inventory Policy," Department of Operations Research Technical Report 104, Stanford (1968).

[10] Palda, K.S., "The Measurement of Cumulative Advertising Effects," Journal of Business *38*, 162-179 (1965).

[11] Pierskalla, W.P., "An Inventory Problem With Obsolescence," Naval Research Logistics Quarterly, *16*, 217-228 (1969).

[12] Shryer, W.A., *Analytical Advertising*, Business Service Corporation, Detroit (1912).

[13] Simon, J.L., "The Effect of Advertising on Liquor Brand Sales," Journal of Market Research *6*, 301-313 (1969).

[14] Simon, J.L., *Issues in the Economics of Advertising*, University of Illinois Press (1970).

[15] Stone, R.F., *Successful Direct Mail Advertising and Selling*, Prentice-Hall, New York (1955).

[16] Telser, L.G., "Advertising and Cigarettes," Journal of Political Economy *70*, 471-499, (1962).

[17] Topkis, D.M., "Minimizing a Submodular Function on a Lattice," Operations Research *26*, 305-21 (1978).

[18] Tull, D.S., "The Carry-Over Effect of Advertising," Journal of Marketing *29*, 46-53 (1965).

[19] Veinott, A.F., Jr., "Optimal Stockage Policies With Nonstationary Stochastic Demand," in H.E. Scarf, D.M. Gilford, and M.W. Shelly (eds.), *Multistage Inventory Models and Techniques*, Stanford University Press, 85-115 (1963).

[20] Veinott, A.F., Jr., "Optimal Policy for a Multiproduct, Dynamic, Nonstationary Inventory Problem," Management Science *12*, 202-222 (1965).

[21] Veinott, A.F., Jr., "Inventory and Production Control," Lecture notes, unpublished (1968).

[22] Veinott, A.F., Jr., "Subadditive Functions on Lattices in Inventory Theory," forthcoming.

[23] Veinott, A.F., Jr., "Monotone Solutions of Problems on Lattice Programming," forthcoming.
[24] Zielske, H., "The Remembering and Forgetting of Advertisement," Journal of Marketing *23*, 239-243 (1959).

APPENDIX

SUBLATTICES AND SUBADDITIVE FUNCTIONS

This whole appendix summarizes some results of Topkis [17] and Veinott [23] (forthcoming) on the minimization of subadditive functions on a sublattice. Throughout, C is the Cartesian product $C_1 \times \ldots \times C_n$ of a finite number of chains. A chain is a completely ordered set. For example, the real line R is a chain.

A subset L of a partially ordered set is called a *lattice* if every two elements s, $s' \in L$ have a greatest lower bound in L, called their *meet* and denoted $s \wedge s'$, and a least upper bound in L, called their *join* and denoted $s \vee s'$.

A subset S of a lattice L is called a *sublattice* of L if S is a lattice and if the meet and join of every two elements of S coincide respectively with the meet and join of those same two points considered as elements of L.

EXAMPLE A1: The set $\{s \in R^n | a_{ij}s_i + b_{ij}s_j \geqslant c_{ij}$, for $1 \leqslant i, j \leqslant n\}$ is a sublattice of R^n provided that $a_{ij}b_{ij} \leqslant 0$ for all $1 \leqslant i, j \leqslant n$.

EXAMPLE A2: A chain in a lattice is a sublattice.

EXAMPLE A3: The Cartesian product of two sublattices is a sublattice.

The set $L_s \equiv \{s' \in L' | (s,s') \in S\}$, where S is a sublattice of the Cartesian product $L \times L'$ of two lattices, is called a *section* of S. The set $\Pi_{L'}S \equiv U_{s \in L} L_s$ is called the *projection* of S on L'.

EXAMPLE A4: The section L_s and the projection $\Pi_{L'}S$ are sublattices of L'.

A real valued function f on a lattice L is called *subadditive* if $f(x \wedge y) + f(x \vee y) \leqslant f(x) + f(y)$ for all $x, y \in L$, and *superadditive* if the inequality is reversed. The assertion that f is subadditive on the product $L \times L'$ of two chains is equivalent to the assertion that the mixed second differences of f are nonpositive on $L \times L'$. If f is twice continuously differentiable and $L \times L'$ is an open subset of R^2, then f is subadditive if and only if $D_{12}f \leqslant 0$ on $L \times L'$.

LEMMA A1: A function f on C is subadditive if and only if $f(s)$ is subadditive in (s_i,s_j) on $C_i \times C_j$ for each $i \leqslant i < j \leqslant n$ and each fixed $s_k \in C_k$ with $k \neq i, j$.

For the next three examples, we define the following functions:

$$f{:}L \times L' \to R,\ g{:}L \to L,\ -g'{:}L' \to L,\ h{:}L' \to L',\ \text{and}\ \hat{g}{:}L \times L' \to L,$$

where L and L' are subsets of R.

EXAMPLE A5: If $f(s,t)$ is subadditive on $L \times L'$ and g and h are nondecreasing on L and L', respectively, then $f(g(s),h(t))$ is subadditive in (s,t) on $L \times L'$.

EXAMPLE A6: If $L + L = L$, $f(s,t)$ is subadditive on $L \times L'$ and convex in s on L for each $t \in L'$, and if g, g' and h are nondecreasing on their respective domains, then $f(g(s) - g'(t),\ h(t))$ is subadditive in (s,t) on $L \times L'$.

EXAMPLE A7: If $f(s,t)$ is subadditive on $L \times L'$ and convex nondecreasing in s on L for each $t \in L'$, $\hat{g}$ is subadditive on $L \times L'$, $\hat{g}$ is nondecreasing on L and nonincreasing on L', and h is nondecreasing on L', then $f(\hat{g}(s,t), h(t))$ is subadditive in (s,t) on $L \times L'$.

EXAMPLE A8: The set of subadditive functions on a lattice is a convex cone.

LEMMA A2: If f is subadditive on C where $C_1 \subset R$, $\bar{F}$ is a chain in the set of all real valued distributions, and $g(t,F) \equiv \int f(s,t)\,dF(s)$ is finite for all $(t,F) \in C_2 \times \cdots \times C_n \times \bar{F}$, then $g(t,F)$ is subadditive in (t,F) on $C_2 \times \cdots \times C_n \times \bar{F}$.

THEOREM A1: (Projection theorem). If S is a nonempty sublattice of the product $L \times L'$ of two lattices, f is subadditive on S, and $g(s) \equiv \inf_{s' \in L_s} f(s,s')$ is finite for $s \in \Pi_L S$, then g is subadditive on $\Pi_L S$.

If $L_1, \ldots, L_n$ are partially ordered sets and $L \subset L_1 \times \cdots \times L_n$, we say $s \in L$ is lexicographically smaller than $t \in L$, written $s \leqslant t$, if either $s = t$ or $s \neq t$ and $s_i \leqslant t_i$ for $i = 1, \ldots, j$, where j is the smallest index i for which $s_i \neq t_i$. If A is a subset of the first n positive integers, denote by L^A the set L where the ordering on L_i is reversed for each $i \in A$. We say $s \in L$ is A-lexicographically smaller than $t \in L$, written $s \leqslant_A t$, if s is lexicographically smaller than t in L^A.

THEOREM A2: (Monotonicity theorem). If S is a nonempty sublattice of the product $L \times L'$ of two lattices where $L' \subseteq R^n$, if L_s is compact for each $s \in \Pi_L S$, and if f is subadditive on S and continuous on L_s for each $s \in \Pi_L S$, then the set $M(s)$ of points minimizing $f(s, \cdot)$ over L_s is a nonempty compact sublattice of L_s for each $s \in \Pi_L S$. Also, $M(s)$ has least, greatest, and A-lexicographically least elements, each of which is nondecreasing in s on $\Pi_L S$.

A DYNAMIC INVENTORY SYSTEM WITH RECYCLING

Morris A. Cohen and William P. Pierskalla*

Department of Decision Sciences
The Wharton School
University of Pennsylvania
Philadelphia, Pennsylvania

Steven Nahmias**

Department of Quantitative Methods
University of Santa Clara
Santa Clara, California

ABSTRACT

This paper deals with a periodic review inventory system in which a constant proportion of stock issued to meet demand each period feeds back into the inventory after a fixed number of periods. Various applications of the model are discussed, including blood bank management and the control of reparable item inventories. We assume that on hand inventory is subject to proportional decay. Demands in successive periods are assumed to be independent identically distributed random variables. The functional equation defining an optimal policy is formulated and a myopic base stock approximation is developed. This myopic policy is shown to be optimal for the case where the feedback delay is equal to one period. Both cost and ordering decision comparisons for optimal and myopic policies are carried out numerically for a delay time of two periods over a wide range of input parameter values.

INTRODUCTION

This paper deals with the analysis of inventory systems in which recycling occurs. The term recycling is used here to indicate that a fixed fraction of the stock used to satisfy demand returns to inventory after a fixed number of periods.

Feedback or recycling in inventory systems can occur in a number of different ways. Examples include systems where customers buy items with a rent/purchase option and return items that are not ultimately purchased. Another cause of recycling is the result of over-ordering stock. This occurs in hospital and regional blood banks since physicians requesting blood for their patients tend to over-order by a factor of two or three. A further application of recycling occurs in retail sales systems where a fixed fraction of stock purchased by customers may be returned, and subsequently mixed with existing inventory.

*Research supported by Grant ENG-77-07463 from the National Science Foundation
**Research supported by Grant 78-3494 from the Air Force Office of Scientific Research and Grant ENG 78-05928 from the National Science Foundation. Visiting, Department of Operations Research Stanford University, 1978-1979

The phenomenon of recycling can also be observed in reparable item inventories. An issue stock inventory is maintained to replace items in the field which are subject to failure. Failed items are returned for repair and after a fixed delay time (which includes the time required for transportation and repair), the repaired item is returned to the issue stock inventory. A fraction of those items that fail are condemned and leave the system forever. The reparable item system is pictured in Figure 1.

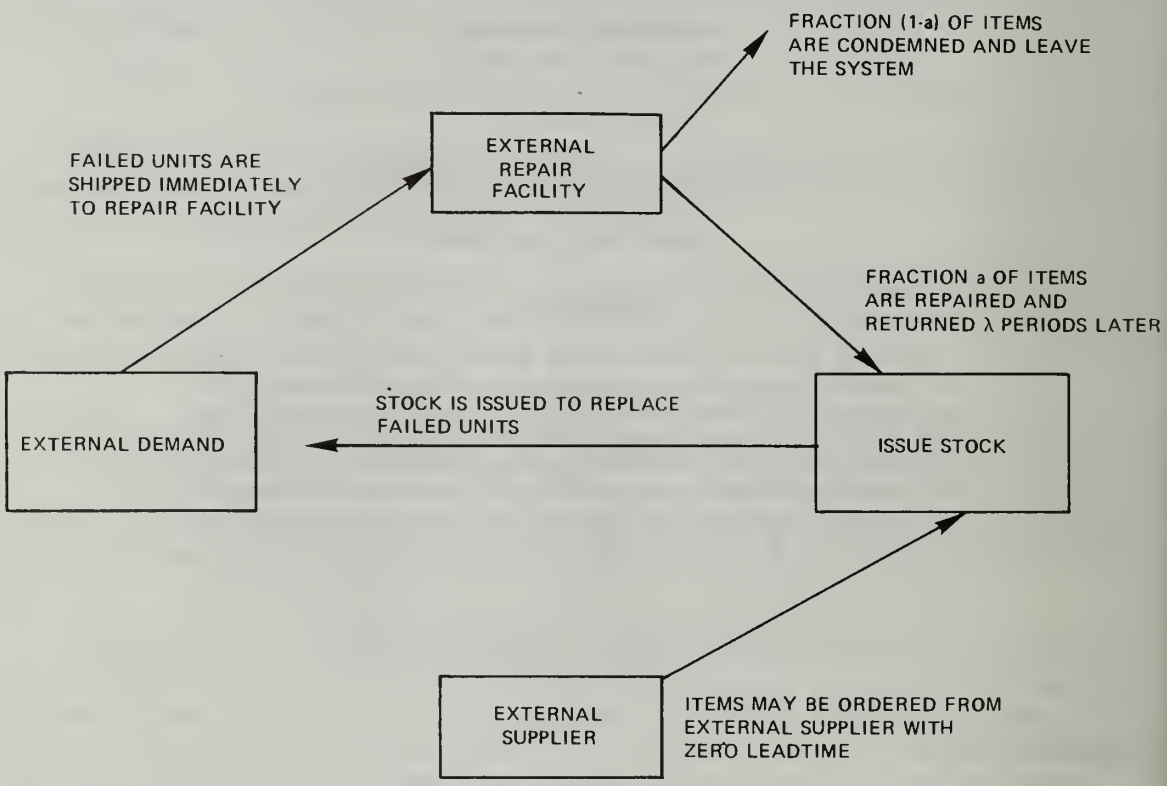

FIGURE 1.

Previous analyses of this class of inventory system have been restricted primarily to simulation studies of blood banks (Cohen and Pierskalla [2]), or systems where demand is deterministic and the proportion of stock recycling is treated as a random variable (Cohen, Nahmias and Pierskalla [3]). Related reparable item inventory models include Prawda and Wright [4] and Allen and D'Esopo [1]. This paper treats the case where demand (or failure) is stochastic and where the fraction of stock that feeds back into the system is fixed.

The paper begins with a description of notation and assumptions for a general model with arbitrary recycle lag and stochastic demand and the functional equation satisfied by the optimal order policy is formulated. A myopic approximation to the optimal order policy and conditions for its optimality are derived for the case of an arbitrary recycle period. The optimality of this policy for the case where the recycle period is equal to one is then demonstrated.

The final section reports on the results of a numerical analysis comparing optimal and myopic policies for a variety of cases where the recycle lag is equal to two periods. These results suggest that the myopic policy provides a very effective approximation to the optimal.

MODEL ASSUMPTIONS AND NOTATION

A periodic review inventory system with the following features is considered:

- Successive demands $\{D_i\}$ are independent and identically distributed random variables with known cumulative distribution function $F(\cdot)$ and density $f(\cdot)$.
- A fixed fraction, a, of stock issued to meet demand is returned after a delay of $\lambda \geqslant 1$ periods. The fraction, $1 - a$, is consumed.
- A fixed fraction, β, of stock on hand at the end of each review period survives, without decay, into the next period and the fraction, $1 - \beta$, is lost to decay.
- Excess demand is lost in each period.
- There is no leadtime for ordering. That is, orders are received in the period in which they are placed.
- Excess demand is lost (lost sales).

Time periods are numbered forward by integers n, $n = 1, 2, \ldots, T$, where T is the decision horizon for the problem.

The following variables describe the state of the system each period:

$\underset{\sim}{I}$ = $(I_1, \ldots, I_{\lambda-1})$ is the vector of stock quantities issued to meet demand in the previous $\lambda-1$ periods. Interpret I_i as the quantity issued exactly i periods previously.

u = starting inventory before ordering but after the arrival of recycled stock in the current period.

z = inventory on hand after ordering and after returns in the current period.

The state of the system at any point in time is described by the vector $(u, \underset{\sim}{I})$, and the decision variable (the order quantity), is given by $z - u$.

We will adopt the common conventions that the holding cost function $h(\cdot)$ and the shortage cost function $p(\cdot)$ are convex functions of ending stock in each period. The outdating cost is assumed to be θ per unit of stock that outdates at the end of each period, and the procurement cost is c per unit. If follows that the one period expected holding, shortage, and outdate cost function, say $L(z)$, is a convex function of the starting stock z, and is given by

$$L(z) = E\{h[(z - D)^+] + p[(D - z)^+] + \theta(1 - \beta)\ (z - D)^+\},$$
$$= \int_0^z h(z - t) f(t)\, dt + \int_z^\infty p(t - z) f(t)\, dt + \theta(1 - \beta) \int_0^z (z - t) f(t)\, dt.$$

Assuming that future costs are discounted by α where $0 < \alpha \leqslant 1$, it follows that the functional equations defining an optimal policy are given by:

$$C_n(u,\underset{\sim}{I}) = \min_{z \geqslant u} \Big\{ c(z-u) + L(z) + \alpha \int_0^z C_{n+1}(\beta(z-t) + aI_{\lambda-1}, t, I_1, \ldots, I_{\lambda-2}) f(t)\,dt + \alpha \int_z^\infty C_{n+1}(aI_{\lambda-1}, z, I_1, \ldots, I_{\lambda-2}) f(t)\,dt \Big\}$$

for $n \geqslant 1$ and $C_{T+1}(\cdot) = 0$. Note that this is equivalent to assuming that stock remaining at the end of the horizon is not salvaged. Other salvage assumptions are possible and will be considered in the next section.

Here $C_n(u,\underset{\sim}{I})$ has the interpretation as the minimum expected discounted cost at the start of period n when u is the starting stock after returns, and $\underset{\sim}{I}$ is the vector of previously issued stock.

The process dynamics are implied in the functional equations above. Let t be the realization of demand in period n. Then there are two cases:

(a) $t \leqslant z$. In this case $(-\beta)(z-t)$ is lost due to decay and $\beta(z-t)$ transfers to the next period which combines with the stock which recycles in period $n+1$, $aI_{\lambda-1}$. Exactly t units are issued to meet the demand. In this case if $(u,\underset{\sim}{I})$ is the state vector in period n, it follows that $(\beta(z-t) + aI_{\lambda-1}, t, I_1, \ldots, I_{\lambda-2})$ is the state vector in period $n+1$.

(b) $t > z$. In this case ending stock in period n is zero and no stock decays or is transferred to the following period. Starting stock the next period consists only of the stock which recycles in period $n+1$, which is $aI_{\lambda-1}$. Since only z can be issued to meet demand, the state vector one period hence is $(aI_{\lambda-1}, z, I_1, \ldots, I_{\lambda-2})$.

The optimal policy is to order the max $(z_n(I) - u, 0)$ where $z_n(\underset{\sim}{I})$, the order to point, minimizes the bracketed term on the right hand side of the functional equation above. Computation of an optimal policy will be difficult due to the limitations of dynamic programming with vector valued state variables. However, for $\lambda = 1$ under reasonably general conditions the optimal policy can be shown to reduce to a single critical number in each period. In addition, a critical number approximation is derived for the case $\lambda > 1$.

A MYOPIC CRITICAL NUMBER APPROXIMATION

We will assume as above, that periods are numbered forwards and the planning horizon is exactly T periods where $\lambda \leqslant T \leqslant +\infty$. We ignore the case $T \leqslant \lambda$, as the feedback process will not be relevant, and the optimal policy reduces to the ordinary critical number order policy. The variables u_n and z_n are still to be interpreted as starting stock after returns before and after ordering respectively in period n. In addition, let $(I(1), \ldots, I(\lambda))$ represent stock issued in the final λ periods. That is, $I(1)$ is issued in period $T-\lambda+1$, $I(2)$ is issued in period $T-\lambda+2$ $\ldots$, and $I(\lambda)$ is issued in period T.

In order to construct the approximation we will need to assume that all stock remaining in the system at the end of the horizon can be salvaged at a return equal to the purchase cost of per unit. This includes stock on hand at that time, (u_{T+1}), and the stock issued in the final λ periods of the horizon $I(1), \ldots, I(\lambda)$. In addition, we assume that the issued stock cannot be salvaged until it returns to inventory. Hence $aI(1)$ is salvaged in period $T+1$, $aI(2)$ in period $T+2, \ldots$, and $aI(\lambda)$ in period $T+\lambda$. (The salvage assumption was first used by Veinott [5].)

With these assumptions, the total discounted cost over T periods, say $TC(\underline{z})$, when following the ordering policy $\underline{z} = (z_1,\ldots,z_T)$ is

$$TC(\underline{z}) = E\left\{\sum_{n=1}^{T} \alpha^{n-1}\{c(z_n - u_n) + L(z_n)\}\right\} - \alpha^T c u_{T+1} - c\sum_{k=1}^{\lambda} \alpha^{T+k-1} a I(k).$$

The process dynamics imply that

$$u_n = \begin{cases} \beta(z_{n-1} - D_{n-1})^+ \text{ for } 2 \leqslant n \leqslant \lambda \\ \beta(z_{n-1} - D_{n-1})^+ + a \min(z_{n-\lambda}, D_{n-\lambda}) \text{ for } \lambda + 1 \leqslant n \leqslant T + 1 \end{cases}$$

and $I(k) = \min(z_{T-\lambda+k}, D_{T-\lambda+k})$ for $1 \leqslant k \leqslant \lambda$.

By a rearrangement of terms, one can show that

$$TC(\underline{z}) = \sum_{n=1}^{T} \alpha^{n-1} W(z_n) - c u_1,$$

where

$$W(z_n) = E[c(z_n - \alpha\beta(z_n - D_n)^+ - \alpha^\lambda a \min(z_n, D_n)] + L(z_n) \text{ for } 1 \leqslant n \leqslant T.$$

We have the following result:

THEOREM 1: Assuming

1. All inventory on hand in periods $T + 1, \ldots, T + \lambda$ can be salvaged in that period at a return of c per unit.
2. $p'(0) > (1 - \alpha^\lambda a)\, c$
3. $P\left\{D_n \geqslant \left(\frac{a + \beta - 1}{\beta}\right) z^*\right\} = 1$ for $1 \leqslant n \leqslant T$
4. $u_1 < z^*$.

where z^* is the minimizing point of $W(z)$ and is the root of the equation $W'(z^*) = 0$, then the optimal policy is to order to z^* every period.

PROOF: Since $\sum_{n=1}^{T} W_n(z^*) \leqslant \sum_{n=1}^{T} W_n(z)$, it follows that z^* is the optimal order to point if z^* can be achieved. It will be possible to order to z^* in period n if and only if $z^* - u_n \geqslant 0$. Following the policy z^* in every period implies that

$$u_n = \beta(z^* - D_{n-1})^+ + a \min(z^*, D_{n-\lambda})$$
$$\leqslant \beta(z^* - D_{n-1})^+ + az^* = \max(az^*, (a + \beta) z^* - \beta D_{n-1}).$$

Clearly $az^* \leqslant z^*$. By assumption 3, $(a + \beta)\, z^* - \beta D_{n-1} \leqslant (a + \beta)\, z^* - (a + \beta - 1)\, z^* = z^*$, hence, $u_n \leqslant z^*$.

Since $W(z)$ is convex in z, and by assumption 2, $W'(0) < 0$, we have that $z^* > 0$.

□

COROLLARY: For $\lambda = 1$, $P\left\{D_n \geqslant \left[\frac{a+\beta-1}{\beta}\right] z^*\right\} = 1$ for $1 \leqslant n \leqslant T$, and hence the myopic policy defined in Theorem 1 is optimal when only assumptions (1), (2) and (4) hold.

PROOF: For $\lambda = 1$ we have,

$$u_n = \beta(z^* - D_{n-1})^+ + a \min(z^*, D_{n-1}),$$

and it follows easily that $u_n \leqslant z^*$ for all realizations of D_{n-1}.

□

When all costs are linear, z^* will be given by,

$$z^* = F^{-1}\left[\frac{p - c(1-\alpha^\lambda a)}{p + h + \theta(1-\beta) - \alpha c(\beta - \alpha^{\lambda-1}a)}\right]. \tag{1}$$

Assumption 3 in Theorem 2 is somewhat tautological when $\lambda \geqslant 2$, since z^* depends on the distribution of D_n. When this assumption does not hold, it may not be possible to order to z^* in every period as it will not necessarily be true that $u_n \leqslant z^*$. However, since the expected cost function, $W(z)$, tends to be relatively flat in a neighborhood of the minimum, it seems reasonable to conjecture that when assumption 3 is not met, ordering $(z^* - u_n)^+$ in every period should give a good approximation.

In order to test this conjecture, numerical computations are performed for $\lambda = 2$ in the next section. Dynamic programming is used to compute the optimal stationary policy which is then compared to z^* for a variety of demand distributions and cost configurations.

NUMERICAL COMPARISONS FOR $\lambda = 2$

A series of runs were carried out for a variety of configurations of the system parameters to compare the effectiveness of the approximation to the optimal policy when the recycle delay was two periods. In order to reduce the number of different factors considered, the cost parameters (c,h,p) are combined into the single constant $m = (p - c)/(p + h)$ (which is motivated by the solution to the newsboy problem). For each demand distribution the following factors and levels are considered:

(1) cost ratio, $m \in \{.5, .75, .95\}$
(2) return fraction, $a \in \{.2, .5, .8\}$
(3) outdate cost, $\theta \in \{1, 2\}$
(4) outdate fraction, $\beta \in \{.2, .8\}$

These factor levels lead to a 36 case experiment. The discount factor α was fixed at .95, order cost c at 1 and holding cost h at .5. The required values of m were achieved by setting p at 2.5, 5.5 and 28.5, respectively.

The total 36 case experiment was run, for uniform, exponential and geometric distributions each with a mean value of five which resulted in a total of 108 cases. The output for a typical case is illustrated in Table I. The optimal solution was computed by standard value iteration techniques and the myopic policy was computed from (1). A 30 period horizon was selected to minimize transient effects. Convergence to the stationary optimal policy generally occurred in ten periods or less. We note, from Table I, that both the optimal solution and cost

penalty for using the myopic approximation are relatively insensitive to changes in the initial inventory level. The maximum percent cost penalty for using the myopic policy is 0.3% in this case.

TABLE I—*Optimal and Approximate Policies for the Case* $m = .5, a = .2, \theta = 1, \beta = .8$, *Poisson Demand*

Initial Inventory I	Optimal Order Function ($z^*(I)$)	Myopic Order-up -to Level (z^*)	Average Cost per Period		% Difference in Cost
			using $z^*(I)$	using z^*	
0	6	5	3.232	3.240	0.2
1	6	5	3.227	3.237	0.3
2	6	5	3.225	3.236	0.3
3	5	5	3.211	3.218	0.2
4	5	5	3.205	3.212	0.2
5	6	5	3.200	3.208	0.3
6	6	5	3.196	3.205	0.3
7	6	5	3.193	1.204	0.3
8	5	5	3.179	3.186	0.2
9	5	5	3.173	3.181	0.3
10	6	5	3.169	3.176	0.2

Table II summarizes 36 runs selected from the set of 108 runs. The runs illustrated were selected by taking the worst case (that is, the largest cost error) over the three demand distributions for each case. Optimal and myopic policies and costs for just the single initial inventory level of five are indicated, since, as noted above, the results are not sensitive to the initial inventory level. Table II also indicates the maximum percent cost differences for each case taken over all initial inventory level values. Note that the maximum percent cost penalty ranges from 0.0% to 6.9% over all factor values. In addition, also note that 71% of all 108 cases had a maximum cost difference of less than 1% and that only 6.5% had a maximum cost difference of more than 5%.

It seems reasonable to conjecture that the myopic policy will also provide a good approximation for values of λ, the recycle delay parameter, larger than two as well. The approximation has the dual advantage of being both easy to compute and easy to implement. The model presented here is applicable to a variety of inventory problems where stock recycling is present, including blood bank inventory control and reparable item management.

ACKNOWLEDGMENT

We would like to acknowledge the assistance of Craig Uthe who wrote the program for the numerical comparisons.

TABLE II— *Worst Case Optimal and Myopic Cost Comparisons*

Run #	*m*	*a*	*theta*	*beta*	Demand CDF	Optimal $z^*(5)$	Optimal Average Cost at $I = 5$	Approx. Policy z^*	Average Cost of Approx. Policy	Maximum % Penalty
1	.50	.2	1	.2	Geometric	2	3.372	1	3.501	4.0
2	.75	.2	1	.2	Poisson	6	4.802	5	5.026	4.7
3	.95	.2	1	.2	Uniform	10	7.658	9	7.969	4.1
4	.50	.5	1	.2	Poisson	5	3.041	4	3.198	5.2
5	.75	.5	1	.2	Uniform	7	5.542	7	5.542	0.0
6	.95	.5	1	.2	Uniform	10	6.890	9	7.21	4.7
7	.50	.8	1	.2	Poisson	5	2.566	4	2.745	6.9
8	.75	.8	1	.2	Uniform	7	4.932	7	4.932	0.0
9	.95	.8	1	.2	Uniform	10	6.188	9	6.525	5.5
10	.50	.2	2	.2	Geometric	1	3.589	1	3.593	.1
11	.75	.2	2	.2	Poisson	5	5.394	5	5.394	0.0
12	.95	.2	2	.2	Poisson	8	7.980	7	8.403	5.3
13	.50	.5	2	.2	Uniform	3	4.649	3	4.649	0.0
14	.75	.5	2	.2	Uniform	7	6.603	6	6.639	0.1
15	.95	.5	2	.2	Poisson	8	7.236	7	7.655	5.9
16	.50	.8	2	.2	Geometric	2	3.095	1	3.188	3.7
17	.75	.8	2	.2	Poisson	6	4.101	5	4.193	2.2
18	.95	.8	2	.2	Poisson	8	6.557	7	7.003	6.7
19	.50	.2	1	.8	Poisson	6	3.200	5	3.208	0.4
20	.75	.2	1	.8	Geometric	6	3.652	5	3.728	2.1
21	.95	.2	1	.8	Geometric	9	4.608	9	4.609	0.0
22	.50	.5	1	.8	Uniform	7	3.161	6	3.217	1.9
23	.75	.5	1	.8	Uniform	9	3.666	8	3.696	0.8
24	.95	.5	1	.8	Poisson	9	3.655	9	3.656	0.0
25	.50	.8	1	.8	Geometric	3	2.162	4	2.185	2.6
26	.75	.8	1	.8	Geometric	6	2.903	6	2.926	1.4
27	.95	.8	1	.8	Geometric	9	3.883	9	3.885	0.2
28	.50	.2	2	.8	Uniform	6	4.090	6	4.093	0.1
29	.75	.2	2	.8	Poisson	7	3.944	6	4.030	2.3
30	.95	.2	2	.8	Poisson	9	4.828	8	5.060	4.8
31	.50	.5	2	.8	Uniform	6	3.417	6	3.427	0.4
32	.75	.5	2	.8	Poisson	7	3.208	6	3.323	3.6
33	.95	.5	2	.8	Poisson	9	4.086	8	4.312	5.5
34	.50	.8	2	.8	Poisson	6	2.070	5	2.092	1.3
35	.75	.8	2	.8	Poisson	7	2.629	6	2.699	2.1
36	.95	.8	2	.8	Poisson	9	3.499	8	3.639	3.9

REFERENCES

[1] Allen, S.G. and D.A. D'Esopo, "An Ordering Policy for Repairable Stock Items," Operations Research, *16*, 669-675 (1968).

[2] Cohen, M.A. and W.P. Pierskalla, "Management Policies for a Regional Blood Bank," Transfusion, *15*, 58-67 (1975).

[3] Cohen, M.A., S. Nahmias and W.P. Pierskalla, "A Feedback Inventory Model of a Hospital Blood Bank," Proceedings of the Seventh Annual Pittsburgh Modelling and Simulation Conference, Instrument Society of America, 612-616 (1976).

[4] Prawda, J. and G.P. Wright, "On a Replacement Problem," Cahiers du Centre d'Etudes de Recherche Operationnelle, *14*, 43-52 (1972).

[5] Veinott, A.F., Jr., "Optimal Policy for a Multi-procduct, Dynamic, Non-stationary Inventory Problem," Management Science, *12*, 206-222 (1965).

SENSITIVITY ANALYSIS AS A MEANS OF REDUCING THE DIMENSIONALITY OF A CERTAIN CLASS OF TRANSPORTATION PROBLEMS

Jacob Intrator

Bar-Ilan University
Ramat-Gan, Israel

Abraham Engelberg

Jerusalem College of Technology
Jerusalem, Israel

ABSTRACT

Sensitivity analysis of the transportation problem is developed in a way which enables reducing the dimensionality of the associated tableau. This technique is used to reduce the dimensionality of a transportation problem whose origin requirements are relatively small at the majority of origins. A long transportation problem, for which efficient solution procedures exist, results. A second application relates to the location-allocation problem. Reducing the dimensionality of such a problem, accompanied by the partial determination of the optimal solution, should prove helpful in the quest for an analytic solution to the aforementioned problem. In the meantime, reducing dimensionality greatly decreases the effort involved in solution by trial and error. Examples of the two applications are provided.

1. INTRODUCTION

The sensitivity analysis of the transportation problem has been thoroughly developed in the classical papers of Srinivasan and Thompson [7]. The latter investigators described how to find the optimal solution of a transformed problem, given the optimal solution of the original problem. Finding the new solution is somewhat easier when the optimum basis structure is preserved. Otherwise, one proceeds from basis to basis until arriving at the optimal basis.

The present paper takes a different approach to sensitivity analysis. Rather than obtaining specific values for all members of the optimal solution, we determine the values of selected variables in a way which enables reducing the dimensionality of the transportation tableau. The greater the similarity between the original and transformed problems, the more the tableau may be reduced in size. Larger transformations do not permit as large a reduction in tableau size, leading to a greater expenditure of time in the solution of the resulting tableau.

The aforementioned technique has been described exhaustively in a paper by Intrator and Paroush [5]. It is not the purpose of the present work to compare these two methods of performing sensitivity analysis as far as efficiency and ease of implementation are concerned, although the authors feel that doing so would indeed be a worthwhile task. Rather, a single facet of the new approach is enlarged upon, and how this aspect may be utilized in speeding up

the solution of two problems not generally associated with sensitivity analysis is explained. More important than the savings in computer time, however, is the fact that reducing the size of the tableau in the case of the location-allocation problem may serve as a stepping-stone in the eventual determination of an analytic solution for the latter problem.

2. DEFINITIONS AND THEOREMS

Consider a transportation problem A having cost matrix $C = (c_{ij})$ and optimal solution (x_{ij}), and a second problem A' identical to the first in all respects except for the fact that the cost matrix has been transformed to $C' = (c'_{ij})$, leading to a new solution (x'_{ij}).

Consider the simple loop

$$L = (i_1,j_1)(i_1,j_2)(i_2,j_2)\ldots(i_k,j_{k+1})$$

for which $j_{k+1} = j_1$. Then the following quantities may be defined:

$$C_L = c_{i_1j_1} - c_{i_1j_2} + c_{i_2j_2} - c_{i_2j_3} - c_{i_kj_{k+1}}, \tag{1}$$

$$C'_L = c'_{i_1j_1} - c'_{i_1j_2} + c'_{i_2j_2} - c'_{i_2j_3} - c'_{i_kj_{k+1}}.\dagger \tag{2}$$

The δL transformation changes any feasible solution of the original problem, (y_{ij}), to a new solution (y^*_{ij}) by defining the members of (y^*_{ij}) as follows:

$$y^*_{ij} = \begin{cases} y_{ij} + \delta, & \text{if } (i,j) = (i_l,j_l), \text{ i.e., if } (i,j) \text{ is an odd-numbered member of } L. \\ y_{ij} - \delta, & \text{if } (i,j) = (i_l,j_{l+1}), \text{ i.e., if } (i,j) \text{ is an even-numbered member of } L. \\ y_{ij}, & \text{if } (i,j) \text{ is not a member of } L. \end{cases} \tag{3}$$

Before executing the δL transformation, both the loop L and the real number δ must be determined. If they are chosen such that (y^*_{ij}) is also a feasible solution, then δL is termed a feasible transformation.

For a given loop L and $\delta < 0$, the δL transformation will be feasible if the odd-numbered cells of the loop are all positive (basic)—a requirement that ensures that $y_{ij} + \delta$ will not become negative. In particular, it is necessary that $-\delta \leqslant \min_{1\leqslant l\leqslant k} y_{i_lj_l}$ (Eq. (3)). If $\delta > 0$, the transformation is feasible if the even numbered cells are positive—a requirement that ensures that $y_{ij} - \delta$ will not become negative. In particular, it is necessary that $\delta \leqslant \min_{1\leqslant l\leqslant k} x_{i_l} y_{i_{l+1}}$. More concisely, a feasible transformation is one for which

$$-\min_{1\leqslant l\leqslant k} y_{i_lj_l} \leqslant \delta \leqslant \min_{1\leqslant l\leqslant k} y_{i_lj_{l+i}}. \tag{4}$$

Letting z and z^* represent the value of the objective function before and after the transformation, respectively, then

†It may be assumed without loss of generality that both C_L and C'_L never equal zero, since it is always possible to appropriately perturb the costs c_{ij} and c'_{ij}.

$$z^* - z = \Sigma\Sigma c_{ij}y_{ij}^* - \Sigma\Sigma c_{ij}y_{ij} = \sum_{1\leqslant l\leqslant k} c_{i_l j_l}(y_{i_l j_l} + \delta)$$

$$+ \sum_{1\leqslant l\leqslant k} c_{i_l j_{l+1}}(y_{i_l j_{l+1}} - \delta) + \sum_{(i,j)\notin L}\sum c_{ij}y_{ij} - \sum_i \sum_j c_{ij}y_{ij}$$

$$= \delta \sum_{1\leqslant l\leqslant k} c_{i_l j_l} - \delta \sum_{1\leqslant l\leqslant k} c_{i_l j_{l+1}} = \delta \cdot C_L. \tag{5}$$

If $\delta C_L < 0$, the objective function has been reduced and δL is called an improving transformation. If $\delta C_L > 0$, δL is called an impairing transformation.

LEMMA 1: The solution (x_{ij}') of the transformed problem A' may be obtained from the solution (x_{ij}) of the original problem A by performing a finite sequence of feasible transformations. The transformations are each improving with respect to A' and impairing with respect to A.

PROOF: If $(x_{ij}) = (x_{ij}')$, no transformations are required, and the theorem is true in a trivial fashion. Assume, on other hand, that $(x_{ij}) \neq (x_{ij}')$. Since $\sum\sum_{ij} x_{ij} = \sum_i a_i = \sum_i a_i'$ (since the only changes are in the cost matrix) $= \sum\sum_{ij} x_{ij}'$, it follows that there exists at least one cell, say (i_1, j_1), such that $x_{i_1 j_1} > x_{i_1 j_1}'$. Since $\sum_j x_{i_1 j} = a_{i_1} = a_{i_1}' = \sum_j x_{i_1 j}'$, there must exist $j_2 \neq j_1$ such that $x_{i_1 j_2}' > x_{i_1 j_2}$. Since $\sum_i x_{ij_2} = b_{j_2} = b_{j_2}' = \sum_i x_{ij_2}'$, there must exist $i_2 \neq i_1$ such that $x_{i_2 j_2} > x_{i_2 j_2}'$. The process continues until a previously utilized row or column is encountered. In this manner, a loop L_1, containing exactly two cells in each row and column, is obtained.

By construction, for each odd-numbered loop member

$$x_{i_l j_l} > x_{i_l j_l}' \rightarrow (x_{i_l j_l}' - x_{i_l j_l}) < 0.$$

For each even-numbered loop member

$$x_{i_l j_{l+1}} < x_{i_l j_{l+1}}' \rightarrow (x_{i_l j_{l+1}} - x_{i_l j_{l+1}}') < 0.$$

In moving from A to A' it is necessary to add to the even-numbered cells of the loop and substract from the odd-numbered cells, an objective which may be achieved by executing a δL transformation with $\delta < 0$. In particular, let

$$\delta_1 = \max_l\{(x_{i_l j_l}' - x_{i_l j_l}), (x_{i_l j_{l+1}} - x_{i_l j_{l+1}}')\}. \tag{6}$$

Then, by Eq. (3), $\delta_1 L_1$ is feasible, because $-\delta_1 \leqslant \min(x_{i_l j_l} - x_{i_l j_l}') \leqslant \min x_{i_l j_l}$.

After the execution of $\delta_1 L_1$, the transformed solution (y_{ij}^*) replaces (x_{ij}) in Eq. (6) and the procedure is repeated, leading to the construction of a second loop L_2. Additional loops L_3, L_4, ... L_k are constructed as necessary. The algorithm terminates when all of the variables have been appropriately modified, i.e., when $\delta_{k+1} = 0$. At each stage, at least one variable attains its optimal value in A', the remaining loop members approach their optimal values in A', while no "overshoots" may occur.

The transformations employed are all feasible with respect to the original solution (x_{ij}) (although they are actually executed on a transformed solution). However, a transformation away from the optimal solution must be an impairing one, implying $\delta_i C_{L_i} > 0$, from which it follows that each transformation is an impairing one with respect to A.

By the same token, the inverse transformations $(-\delta_i)L_i$ gradually lead from A' to A when executed in the sequence $(-\delta_k)L_k$, $(-\delta_{k-1})L_{k-1}$, ... $(-\delta_1)L_1$. Since (x'_{ij}) is optimal, these transformations are impairing with respect to A', implying that the original transformations are improving with respect to A'.

COROLLARY: If $c'_{ij} = c_{ij}$ in all but one row, say i_0, then each of the loops L_i must contain members of row i_0.

The corollary follows from the fact that the transformation loops are improving ones with respect to A', i.e., $\delta_i C'_{L_i} < 0$, and impairing ones with respect to A, i.e., $\delta_i C_{L_i} > 0$, which can only occur if $C_{L_i} \neq C'_{L_i}$, i.e., the loops contains members of i_0.

DEFINITION: Let $S_i = \sum_{\substack{1 \leqslant j \leqslant n \\ x_{ij} > x'_{ij}}} (x_{ij} - x'_{ij})$, i.e., S_i represents the sum of the positive differences between the optimal values of the variables in row i of the original problem A and the transformed problem A'.

Let $S = \max_i S_i$, i.e., S represents the maximum such difference among all the rows.

LEMMA 2: If $c_{ij} = c'_{ij}$ in all but one row, say i_0, then $S = S_{i_0} \leqslant a_{i_0}$.

PROOF: By the Corollary to Lemma 1, every loop L_i contains members of row i_0. But row i_0 contains initially only a_{i_0} items, hence $S_{i_0} = \sum_{i=1}^{k} \delta_i \leqslant a_{i_0}$. As far as the other rows are concerned, $p \leqslant k$ of the loops L_i pass through any specific row, so that

$$S_i = \sum_{i=1}^{p} \delta_i \leqslant \sum_{i=1}^{k} \delta_i \leqslant a_{i_0}.$$

Thus, $S = \max S_i \leqslant a_{i_0}$.

It will now be shown that the case in which A and A' differ in terms of origin requirements reduces to the case in which the two differ in terms of transportation costs.

THEOREM 1: Suppose the two transportation problems A and A' differ only with respect to their origin requirements. In particular, suppose that

(a) $a'_i \leqslant a_i$ $(i = 1,. . .m - 1)$

(b) $a'_m > a_m$

(c) $\sum_{i=1}^{m} a'_i = \sum_{i=1}^{m} a_i$.

Then $x'_{mj} \geqslant x_{mj}$, where (x_{ij}) and (x'_{ij}) represent the optimal solutions of A and A' respectively.

PROOF: Suppose, on the contrary, that there exists some column j, such that $x_{mj_1} > x'_{mj_1}$. Since $b_{j_1} = \sum_{i=1}^{m} x_{ij_1} = \sum_{i=1}^{m} x'_{ij_1}$, there must exist at least one $i_2 \neq i_1 = m$ such that $x'_{i_2 j_1} > x_{i_2 j_1}$. However, from

$$a'_{i_2} = \sum_{j=1}^{n} x'_{ij_2} \leqslant a_{i_2} = \sum_{j=1}^{n} x_{i_2 j},$$

it follows that there must be some $j_2 \neq j_1$ such that $x_{i_2 j_2} > x'_{i_2 j_2}$. For the new j_2 there must be a new i_3. The process continues until one returns to a row or column encountered previously in the process, and a loop, say L, containing k cells is obtained.

Regarding loop L there exist three possibilities:

(a) $C_L = 0$. Perturbation eliminates this possibility.

(b) $C_L > 0$. Let $\delta = \min_{1 \leqslant l \leqslant k} x_{i_l j_l} > \min_{1 \leqslant l \leqslant k} x'_{i_l j_l} \geqslant 0$. Then the $(-\delta)L$ transformation performed on A is an improving one, since $-\delta C_L < 0$, a result which contradicts the optimality of (x_{ij}).

(c) $C_L < 0$. Let $\delta = \min_{1 \leqslant l \leqslant k} x'_{i_l j_{l+1}} > \min_{1 \leqslant l \leqslant k} x_{i_l j_{l+1}} \geqslant 0$. Then the δL transformation performed on A' is an improving one, since $\delta C_L < 0$, a result which contradicts the optimality of (x'_{ij}).

The initial assumption that there exists a variable $x_{mj_1} > x'_{mj_1}$ must therefore be dropped in favor of the conclusion that $x_{mj_1} \leqslant x'_{mj_1}$. (Note that it follows from Theorem 1 that $x'_{mj} = 0$ implies $x_{mj} = 0$ while $x_{mj} > 0$ implies $x'_{mj} > 0$ $(1 \leqslant j \leqslant n)$ when the stipulated conditions hold.)

THEOREM 2: Suppose that A and A' differ in the following respects as regards their origin requirements:

$$a'_1 = a_1 + \alpha,\ a'_m = a_m - \alpha,\ a'_i = a_i (i = 2, m-1).$$

Then $S_1 = 0$, $S_m = \alpha$, $S_i \leqslant \alpha$, $2 \leqslant i \leqslant m-1$, i.e., $S = \alpha$.

PROOF: By Theorem 1, considering a_1 to be a_m, it follows that $x'_{1j} > x_{1j} (1 \leqslant j \leqslant n)$, which implies that $S_1 = \sum_{x_{1j} > x'_{1j}} (x_{1j} - x'_{1j}) = 0$.

Also, looking upon the new problem as the original and vice-versa, it follows from Theorem 1 that $x'_{mj} \leqslant x_{mj}$, which implies that

$$S_m = \sum_{\substack{j=1 \\ x_{mj} > x'_{mj}}}^{n} (x_{mj} - x'_{mj}) = a_m - a'_m = \alpha.$$

In order to prove that $S_i \leqslant \alpha (1 < i < m)$, it will be demonstrated that A and A' are equivalent to two transportation problems which differ from each other with respect to the transportation costs in a single row exclusively.

Consider the auxiliary $(m+1) \times n$ transportation problem $\hat{A}$ having the following parameters:

$$\hat{a}_i = a_i,\ i = 1, 2, \ldots m-1,\ \hat{a}_m = a_m - \alpha,$$

$$\hat{a}_{m+1} = \alpha,$$

$$\hat{c}_{ij} = c_{ij},\ \hat{b}_j = b_j\ (1 \leqslant i \leqslant m,\ 1 \leqslant j \leqslant n),$$

$$\hat{c}_{m+1,j} = c_{mj},\ 1 \leqslant j \leqslant n.$$

Let $\hat{A}'$ be defined identically, except that $\hat{c}_{m+1,j} = c_{1j}$, $1 \leqslant j \leqslant n$. Now, note that $\hat{A}$ is equivalent to the original problem A with $x_{mj} = \hat{x}_{mj} + \hat{x}_{m+1,j}$, since $\Sigma(\hat{x}_{mj} + \hat{x}_{m+1,j}) = (a_m - \alpha) + \alpha = a_m$ (see [3], pp. 319-320). On the other hand, $\hat{A}'$ is equivalent to A' with $x'_{1j} = \hat{x}_{1j} + \hat{x}_{m+1,j}$, since $\Sigma(\hat{x}_{1j} + \hat{x}_{m+1,j}) = a_1 + \alpha$. Since $\hat{A}$ and $\hat{A}'$ are identical except for differing transportation costs in row $(m+1)$ for which $a_{m+1} = \alpha$, it follows from Lemma 2 that $S_i \leqslant S = S_{m+1} \leqslant \alpha$.

THEOREM 3: Suppose A and A' differ as follows with respect to their origin requirements:

(a) $a'_p = a_p - \gamma_p$ $(1 \leqslant p \leqslant r,\ 0 \leqslant \gamma_p \leqslant a_p)$,

(b) $a'_s = a_s + \beta_s$ $(r < s \leqslant m,\ \beta_s \geqslant 0)$,

(c) $\Sigma\beta_s = \Sigma\gamma_p$.

Then

$$S = \Sigma\beta_s = \sum_{a'_s > a_s} (a'_s - a_s)$$

and

$$S_i \leqslant \sum_{s \neq i} \beta_s = \Sigma\beta_s - \beta_i = \sum_{a'_s > a_s} (a'_s - a_s).$$

PROOF: The transition from A to A' may be broken down into a finite sequence of steps of the type considered in Theorem 2. The process may be described as follows. Suppose $\min\{\gamma_p, \gamma_p > 0\} = \gamma_{p_1}$ and $\min\{\beta_s, \beta_s > 0\} = \beta_{s_1}$. Let $\alpha_1 = \min(\gamma_{p_1}, \beta_{s_1})$. Then consider the transformation $a^{(1)}_{p_1} = a_{p_1} - \alpha_1$, $a^{(1)}_{s_1} = a_{s_1} + \alpha_1$. By Theorem 2,

$$S_i = 0,\ i = s_1,$$

$$S_i = \alpha_1,\ i = p_1,$$

$$S_i \leqslant \alpha_1,\ i \neq p_1,\ s_1,$$

$$S = \alpha_1.$$

Note that either $a^{(1)}_{s_1} = a'_{s_1}$ or $a^{(1)}_{p_1} = a'_{p_1}$ (or both). The remaining values of a_i are unaffected. In other words, one of the requested changes has been executed, a second has been (at least) partially implemented, and no "overshoots" have occurred.

The second step involves choosing the next origin requirement to be transformed to a'_i. First, one sets $\gamma_{p_1} = \gamma_{p_1} - \alpha_1$ and $\beta_{s_1} = \beta_{s_1} - \alpha_1$ (at least one of which will be 0), and then the previously described algorithm is executed. The process terminates after k steps when all γ_p and β_s equal zero. Applying Theorem 2 to each of the k stages, one finds that

$$S = \sum_{h=1}^{k} \alpha_h = \sum_{s=r+1}^{m} \beta_s$$

(since each β_s has been reduced to 0). For $r < i \leqslant m$, $S_i = 0$ for every transformation which leads to an increase in the value of a_i. Such transformations involve a total increase of β_i. Taking all transformations into account (even those which do not involve a_i, for which $S_i \leqslant \alpha_h$), one finds that $S_i \leqslant \sum_{h=1}^{k} \alpha_h - \beta_i = \sum_{s=r+1}^{m} \beta_s - \beta_i = \sum_{\substack{s=r+1 \\ s \neq i}}^{m} \beta_s$. The latter applies when $i \leqslant r$ as well, since s cannot equal i in such a case.

DEFINITION: Let $D_{ik}(j) = c_{ij} - c_{kj}$, i.e., $D_{ik}(j)$ equals the difference in costs between the cells of column j associated with rows i and k, respectively. Since it is always possible to introduce cost perturbation, it may be assumed that $D_{ik}(j_1) = D_{ik}(j_2)$ if and only if $j_1 = j_2$.

LEMMA 3: If $x_{k\nu} > 0$ and $x_{i\mu} > 0$, then $D_{ik}(\nu) > D_{ik}(\mu)$, where (x_{ij}) represents the optimal solution of transportation problem A.

PROOF: Consider the simple loop $L = (k,\mu); (k,\nu); (i,\nu); (i,\mu)$. Letting δ equal the minimum of the even-numbered cells (i.e., $\delta = \min(x_{k\nu}, x_{i\mu}) > 0$ by hypothesis) ensures that the δL transformation is feasible. However, (x_{ij}) is the optimal solution, so that δL cannot be an improving transformation; i.e., of necessity $\delta C_L = \delta(c_{k\mu} - c_{k\nu} + c_{i\nu} - c_{i\mu}) = \delta[D_{ik}(\nu) - D_{ik}(\mu)] > 0$, which implies that $D_{ik}(\nu) > D_{ik}(\mu)$, as was to be proven. See Fig. 1 for a schematic representation.

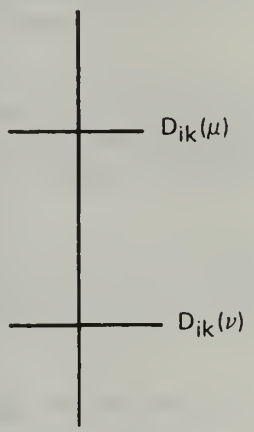

FIGURE 1. The relative positions of $D_{ik}(\mu)$ and $D_{ik}(\nu)$ on a scale whose uppermost element corresponds to the smallest value of $D_{ik}(j)$ and whose lowermost element corresponds to the largest value of $D_{ik}(j)$ when it is known that $x_{k\nu} > 0$ and $x_{i\mu} > 0$.

COROLLARY: If $D_{ik}(\nu) < D_{ik}(\mu)$, it is impossible that both $x_{k\nu} > 0$ and $x_{i\mu} > 0$.

DEFINITION: Let A_i represent the set of columns associated with the optimal group of basic cells in row i of problem A, i.e., $A_i = \{j | x_{ij} > 0\}$. The n_i elements of A_i may be arranged as the vector $K_{ik} = (i_1, \ldots i_{n_i})$ where $D_{ik}(i_1) < D_{ik}(i_2) < \ldots < D_{ik}(i_{n_i})$, i.e., each basic column j is ordered in terms of its corresponding $D_{ik}(j)$ value. The integer $s(i,k)$ will be determined by the following inequality:

$$\sum_{\mu=s(i,k)}^{n_i} x_{ii_\mu} > S_i \geqslant \sum_{\mu=s(i,k)+1}^{n_i} x_{ii_\mu}. \tag{7}$$

For an example of the above notation, see Fig. 2.

Consider the transportation problem A' which differs from A with respect to origin requirements exclusively. Although the optimal solution (x'_{ij}) differs from (x_{ij}), the $D_{ik}(j)$ values are unaltered, since the transportation costs have not changed.

THEOREM 4: If $\nu < s(i,k)$, then $x'_{ki_\nu} = 0 (i_\nu \in K_{ik})$.

PROOF: Suppose $x'_{ki_\nu} > 0$. Then, by the Corollary to Lemma 3, for all $\nu < \mu \leqslant n_i$ (for which $D_{ik}(i_\nu) < D_{ik}(i_\mu)$), $x'_{ii_\mu} = 0$. Since $i_\mu \in A_i$, $x_{ii_\mu} > 0$. It follows, then, that $x_{ii_\mu} > x'_{ii_\mu}$ for all $\nu < s(i,k) \leqslant \mu \leqslant n_i$. Then

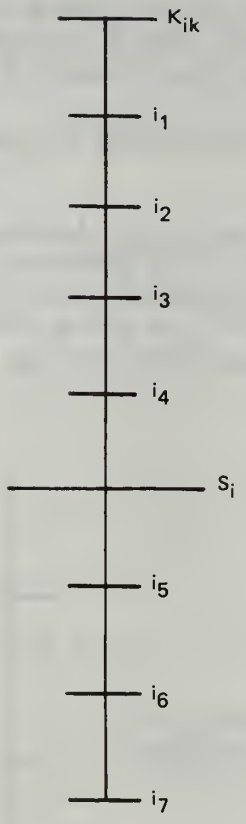

FIGURE 2. In the above figure, $s(i,k) = 4$ and $n_i = 7$, where $i_1, i_2, \ldots i_7$ are the elements of A_i.

$$\sum_{\mu=s(i,k)}^{n_i} x_{ii_\mu} = \sum_{\mu=s(i,k)}^{n_i} (x_{ii_\mu} - x'_{ii_\mu}) \leqslant \sum_{x_{ij}>x'_{ij}} (x_{ij} - x'_{ij}) = S_i.$$

In short, $\sum_{\mu=s(i,k)}^{n_i} x_{ii_\mu} \leqslant S_i$. But by definition $\sum_{\mu=s(i,k)}^{n_i} x_{ii_\mu} > S_i$, which is in contradiction to the result presently obtained. Accordingly, the initial assumption that $x'_{ki_\nu} > 0$ must be rejected.

The application of Theorems 3 and 4 is as follows. One first calculates $s(i,k)$ for each of the $m(m-1)$ permutations of i and k, a task which requires knowledge concerning S_i, $i = 1, \ldots m$. Theorem 3 enables the determination of an upper bound on S_i, which in turn provides a lower bound on $s(i,k)$, as follows from Eq. (7). Utilizing Theorem 4, a large number of variables x'_{kj} are set to 0. Whenever all but one of the variables in a given row or column equal 0, the remaining variable equals a_i or b_j as the case may be, and the respective row or column may be eliminated from the tableau.

3. APPLICATIONS

Reducing the Dimensionality of a Certain Class of Transportation Problems

Consider an $m \times n$ transportation problem A for which $\sum_{i=1}^{m} a_i = \sum_{j=1} b_j = R$. Furthermore, suppose that $\sum_{i=1}^{m_1} a_i = \alpha R$ and $\sum_{i=m_1+1}^{m} a_i = (1-\alpha)R$. The present application concerns the case in which $m_1 < < m$ and α is near 1, i.e., the situation in which most of the merchandise is

concentrated at a relatively small number of warehouses. This would be the case when a small number of factories, at which large quantities may be stored, supply a larger number of intermediary warehouses of significantly smaller storage capacity.

Consider the following three transportation problems:

A: The original problem,

A': An $m_1 \times n$ transportation problem identical to A except for the fact that the last $m - m_1$ rows of A are deleted and $a'_{m_1} = \sum_{i=m_1}^{m} a_i$,

A'': An $m \times n$ transportation problem identical to A except for the fact that the last $m - m_1$ values of a_i are set to 0 and $a'_{m_1} = \sum_{i=m_1}^{m} a_i$. Clearly, an optimal solution A' also optimizes A''. The proposed solution procedure is as follows:

(a) Solve A'. The relatively small value of m_1 ensures that A' will be a long transportation problem for which an extremely efficient computerized solution procedure has been developed [4]. The solution to problem A'' is now available, and the previously obtained sensitivity analysis results will now be utilized in finding the solution to A.

(b) Based on the optimal solution to A'', find the set of positive cells in row i, A_i, which serves as a prerequisite in determining the ordered vector K_{ik} for each value of i, $1 \leqslant i \leqslant m_1$ and $1 \leqslant k \leqslant m$, $k \neq i$, leading to a total of $m_1(m-1)$ vectors K_{ik}.

(c) For each i, $1 \leqslant i \leqslant m_1$, determine S_i. Note that A and A'' differ only with respect to their requirements a_i. More specifically,

$$a_i = a_i'', \quad i = 1, \ldots m_1 - 1,$$

$$a_{m_1} = a''_{m_1} - \gamma_{m_1}$$

(where $\gamma_{m_1} = \sum_{i=m_1+1}^{m} a_i$, since $a''_{m_1} = \sum_{i=m_1}^{m} a_i$),

$$a_i = a_i'' + \beta_i, \quad i = m_1 + 1, \ldots m \text{ (where } \beta_i = a_i, \text{ since } a_i'' = 0 \text{ over this range)}.$$

Since $\gamma_{m_1} = \sum_{i=m_1+1}^{m} \beta_i$, Theorem 3 is applicable, implying that

$$S_i \leqslant \sum_{i=m_1+1}^{m} (a_i - a_i'') = (1-\alpha)R, \quad i = 1, \ldots m_1.$$

(d) For each $1 \leqslant i \leqslant m_1$ and $1 \leqslant k \leqslant m$, $k \neq i$, determine $s(i,k)$ by summing the values of x_{ij}'' associated with the lower values of the ladder scale K_{ik} until they exceed $(1-\alpha)R$ (see Eq. 7).

(e) By Theorem 4, for $\nu < s(i,k)$, $x_{ki_\nu} = 0 (i_\nu \in K_{ik})$. Since $\sum_{\mu=1}^{n_i} x_{ii_\mu} = a_i$, the theorem will be useful if $s(i,k) > 1$, i.e., if $\sum_{\mu=s(i,k)}^{n_i} x_{ii_\mu} < a_i$, which will imply that

$x_{ki_1}, x_{ki_2}, \ldots x_{k[s(i,k)-1]} = 0$. But in the present situation $\sum_{\mu=s(i,k)}^{n_i} x_{ii_\mu} \cong S_i \leqslant (1-\alpha)R$ which by assumption is small with respect to a_i, so that $s(i,k) > > 1$ and hence a large number of variables x_{kj} will be set to zero in row k.

(f) The $m_1(m-1)\ s(i,k)$ values each lead to a set of x_{kj} values which are known to equal zero in the original problem A. Since $1 \leqslant k \leqslant m$, such zeroes have been determined for all of the m rows of A. Whenever all but one of the variables in a given row or column equal zero, the remaining variable equals a_i or b_j as the case may be and the respective row or column may be eliminated from the tableau.

(g) A second means of obtaining zeroes is available for row m_1 exclusively. Since $a_i'' \leqslant a_i$ $(i = 1,\ldots m, i \neq m_1)$ and $a_{m_1}'' = \sum_{i=m_1}^{m} a_i > a_{m_1}$, it follows from Theorem 1 that $x_{m_1 j}'' > x_{m_1 j}$. Then if $x_{m_1 j}'' = 0$, it follows that $x_{m_1 j} = 0$ in problem A.

(h) Steps (f) and (g) serve to reduce the row and column dimensions of the remaining problem. The reduced tableau A_1 may be treated in two ways:

- It may be solved as a routine transportation problem,
- It may be solved using the technique of Srinivasan and Thompson [7]. If A'' is reduced in size in a manner which completely parallels the reduction in size of the original problem A, leading to A_1'', then an optimal solution to A_1'' is available, since an optimal solution to the original problem A_1'' is optimal for any subtableau.

Selecting Additional Warehouses from a Predetermined Set of Possibilities

Consider the following problem. Given

(a) The requirements at each of n destinations,

(b) The capacity at each of m_1 origins,

(c) The transportation costs between each of the m_1 origins and each of the destinations,

(d) The transportation costs between each of $(m - m_1)$ additional (potential) origins and each of the given destinations,

(e) The total additional capacity desired, i.e., $\sum_{i=m_1+1}^{m} a_i$,

Find

- The specific values of a_i, $m_1 + 1 \leqslant i \leqslant m$,
- The amount to be shipped from each origin to each destination so as to minimize the total shipping cost. The number of origins in the optimal solution equals the number of sources i for which $x_{ij} > 0$ for at least one value of j. It is expected that some of the a_i, $m_1 + 1 \leqslant i \leqslant m$, will equal zero, i.e., some of the warehouses will not be built.

The problem described is reminiscent of the location-allocation problem [1,2] in the sense that the capacities and final locations of the additional origins are not fixed in advance. However, rather than choosing coordinates optimally, the optimal capacities and locations are indicated by the values of a_i which are positive for minimal cost schemes. The final locations are thus limited to being a subset of the original set of potential warehouses of size $m - m_1$.

As in the first application, the procedure to be presented relates to the situation in which

(a) $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j = R$,

(b) $\sum_{i=1}^{m_1} a_i = \alpha R$, $\sum_{i=m_1+1}^{m} a_i = (1-\alpha)R$,

(c) $m_1 < < m$,

(d) α is close to 1.

The first m_1 origins are assumed to be operating and indeed supplying most of the merchandise, since $\sum_{i=1}^{m_1} a_i = \alpha R$. Accordingly, $a_1, a_2, \ldots, a_{m_1}$ will be assumed to be known. The question relates only to the exact and optimal values of $a_{m_1}, a_{m_1+1}, \ldots a_m$.

The approach to be taken involves the solution of a number of long transportation problems. In particular, consider the following transportation problems:

- A, an $m \times n$ transportation problem, with c_{ij} equal to the known transportation cost between each of the m origins and n destinations. A cannot be solved initially because no specific values have been specified for $a_{m_1+1}, \ldots a_m$.

- A'_{m_1+1}, an $(m_1 + 1) \times n$ transportation problem for which

$$c'_{ij} = c_{ij},\ i = 1, \ldots m,\ j = 1, \ldots n,$$
$$c'_{m_1+1,j} = c_{m_1+1,j},\ j = 1, \ldots n,$$
$$a'_i = a_i,\ i = 1, \ldots m_1,$$
$$a'_{m_1+1} = \sum_{i=m_1}^{m} a_i = (1-\alpha)R.$$

- A'', an $m \times n$ transportation problem identical to A except for the fact that the last $m - (m_1 + 1)$ values of a_i are set to 0, and $a''_{m_1+1} = \sum_{i=m_1}^{m} a_i$.

Clearly, any optimal solution A'_{m_1+1} also optimizes A''.

A'_{m_1+1} is a long transportation problem which can be solved efficiently as previously noted, thus providing a solution to A''. But the relationship between A and A'' may be stated as follows:

$$a''_i \leqslant a_i,\ i=1, \ldots m,\ i \neq m_1 + 1,$$
$$a''_{m_1+1} = \sum_{i=m_1+1}^{m} a_i > a_{m_1+1}.$$

It follows then, from Theorem 1, that

(i) $x''_{m_1+1,j} \geqslant x_{m_1+1,j}$

and from the note following Theorem 1 that

(ii) $x''_{m_1+1,j} = 0$ implies $x_{m_1+1,j} = 0$ for all $1 \leqslant j \leqslant n$.

A method for determining zeroes in row $m_1 + 1$ of the original problem has been presented. The number of zeroes determined will be quite large, since $\sum_{j=1}^{n} x''_{m_1+1,j} = a''_{m_1+1} = (1 - \alpha)R$ has been stipulated to be small.

In general, since the fraction of the total requirements assigned to row m_1 equals $\frac{(1-\alpha)R}{R}$, the average number of zeroes in row m_1 will be αn, where α is close to one.

Zeroes may be obtained in rows $m_1 + 2$, $m_1 + 3$, ... m by following an identical procedure. For example, A'_{m_1+2} would also be an $(m_1 + 1) \times n$ problem which differs from A'_{m_1+1} only regarding the transportation costs in row $m_1 + 1$, which would now equal $c_{m_1+2,j}$. In general, for A'_{m_1+k}, $k = 1, \ldots, m - m_1$,

$$c_{m_1+1,j} = c_{m_1+k,j} \quad j = 1, \ldots n.$$

Zeroes have thus been determined for rows $m_1 + 1, \ldots m$. As far as the first m_1 rows are concerned, a similar procedures may be followed, although it will be a little less fruitful. Again the total amount of merchandise associated with the lower rows, $(1 - \alpha)R$, will be combined. However, instead of letting $a'_i = (1 - \alpha)R$ for one of the lower rows, $(1 - \alpha)R$ will be added to one of the first m_1 rows, and a_i will be set to zero for $i = m_1 + 1, \ldots m$. The expected number of zeroes in row $i = 1, \ldots m_1$ will then be $\left(\frac{\alpha R - a_i}{R}\right) \cdot n = \left(\alpha - \frac{a_i}{R}\right) \cdot n$.

As in the previous application, the large number of zeroes obtained enables the elimination of numerous rows and columns of the original tableau A, and the subsequent reduction in size of the remainder of the problem. To solve the reduced problem, one must supply values for $a_{m_1+1}, \ldots a_m$. However, at this point, the dimensions of the problem to be solved are sufficiently small that a large number of alternative sets of values for $\{a_{m_1+1} \ldots a_m\}$ may be experimented with until a satisfactory solution is obtained.

4. AN EXAMPLE

The utility of the preceding algorithms can be demonstrated only by systematically experimenting on large-scale problems. The purpose of the present small-scale example, on the other hand, is to review and illustrate the previously described techniques, terminology, and notation. Consider the 5×10 transportation problem which appears as Table 1. Note that most of the merchandise is concentrated at origin 1, so that $m_1 = 1$.

A' will be the 1×10 transportation problem with $a_{m_1} = \sum_{i=m_1}^{m} a_i = 59$, and the optimal solution to A' appears in Table 2.

The set A_1 includes all columns of A' having a basic variable in row 1. In the present example, $A_1 = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. For $j \in A_1$, the value of $D_{12}(j)$ will now be determined (Table 3) and arranged in ascending order (Table 4).

TABLE 1

b_j / a_i	5	8	8	3	4	9	2	9	8	3
51	4	5	7	1	6	7	3	2	1	4
1	2	1	5	3	8	1	4	9	2	1
2	3	8	1	5	4	2	9	3	1	2
2	0	5	8	2	3	5	9	4	2	6
3	2	0	2	1	5	2	2	4	3	2

TABLE 2

b_j / a_i	5	8	8	3	4	9	2	9	8	3
59	4	5	7	1	6	7	3	2	1	4
	5	8	8	3	4	9	2	9	8	3

TABLE 3

Destination j	1	2	3	4	5	6	7	8	9	10
$D_{12}(j)$	2	4	2	−2	−2	6	−1	−7	−1	3

TABLE 4

Destination j	8	4	5	7	9	1	3	10	2	6
$D_{12}(j)$	−7	−2	−2	−1	−1	2	2	3	4	6
x_{ij}	9	3	4	2	8	5	8	3	8	9

$K_{1k} = \{8, 4, 5, 7, 9, 1, 3, 10, 2, 6\}$.

$$S_1 \leqslant \sum_{i=1}^{m} (a_i - a_i'') = 59 - 51 = 8$$

$$\sum_{\mu=10}^{10} x_{ii_\mu} = 9 > S_1 = 8.$$

Thus $s(1, 2) = 10$. It follows that $x'_{2j} = 0 (1 \leqslant j \leqslant 10,\ j \neq 6)$.

Similarly, if the $D_{13}(j)$ are arranged in an increasing sequence one obtains Table 5.

TABLE 5

Destination j	7	4	2	8	9	1	10	5	6	3
$D_{13}(j)$	−6	−4	−3	−1	0	1	2	2	5	6
x_{ij}	2	3	8	9	8	5	3	4	9	8

$$\sum_{\mu=9}^{10} x_{ii_\mu} = 17 > S_1 \leqslant 8$$

Thus $x'_{3j} = 0$ for $j = 1, 2, 4, 5, 7, 8, 9$.

For $D_{14}(j)$ and $D_{15}(j)$, the results of Tables 6 and 7 are obtained.

TABLE 6

j	7	8	10	3	4	9	2	6	5	1
x_{ij}	2	9	3	8	3	8	8	9	4	5

Thus $x'_{4j} = 0$ for $j = 2, 3, 4, 6, 7, 8, 9, 10$.

TABLE 7

j	8	9	4	5	7	1	10	6	3	2
x_{ij}	9	8	3	4	2	5	3	9	8	8

$x'_{5j} = 0$ for $j = 1, 4, 5, 6, 7, 8, 9, 10$.

The results of the present calculations are arranged in a 5×10 table (Table 8) enabling the reduction of the original problem A to the tableau of Table 9. Note that the second means of obtaining zeroes suggested in the previous section is not applicable here since the variables in row $m_1 = 1$ are all positive.

TABLE 8

i \ j	1	2	3	4	5	6	7	8	9	10	a_i
1				3			2	9	8	3	51
2	x	x	x	x	x	1	x	x	x	x	1
3	x	x		x	x		x	x	x	x	2
4		x	x	x		x	x	x	x	x	2
5	x			x	x	x	x	x	x	x	3
b_j	5	8	8	3	4	9	2	9	8	3	59

TABLE 9

	1	2	3	5	6	b_j
1						51 − 25 = 26
3	x	x		x		2
4		x	x		x	2
5	x			x	x	3
a_i	5	8	8	4	8	

In Tables 8 and 9, an x in location (i,j) means that variable (i,j) has been determined to be equal to zero. Note that the original values of a_i and b_j have been reduced by the optimal values of x_{ij} which have already been determined; e.g., instead of $a_1 = 51$, $a_1 - x_{14} - x_{17} - x_{18} - x_{19} = 26$ appears in Table 9. Similarly, b_6 has been reduced by one. The reduced problem is of dimension 4×5 rather than 5×10.

The sample problem previously described will be used to demonstrate the second application. As before, $a_1 = 51$. Suppose $\sum_{i=2}^{5} a_i = 8$ as before, but the exact values of these origin requirements are unknown, and it is desired to determine them optimally.

In the present case, the number of known values of a_i, i.e. m_1, equals one. $A'_{m_1+1} = A'_2$ is formed by setting $a'_2 = \sum_{i=2}^{5} a_i = 8$, and appears in Table 10 together with its optimal solution.

TABLE 10

b_j / a_i	5	8	8	3	4	9	2	9	8	3
51	4 / 5	5 / 8	7 / 8	1 / 3	6 / 4	7 / 4	3 / 2	2 / 9	1 / 8	4
8	2	1	5	3	8	1 / 5	4	9	2	1 / 3

$A'_{m_1+2} = A'_3$ is formed by setting $a'_3 = \sum_{i=2}^{5} a_i = 8$, and appears in Table 11 together with its optimal solution, as do A'_4 (Table 12) and A'_5 (Table 13). It should be noted that the same initial solution may be used when solving each of the auxiliary problems.

TABLE 11

b_j / a_i	5	8	8	3	4	9	2	9	8	3
51	4 / 5	5 / 8	7 / 3	1 / 3	6 / 4	7 / 9	3 / 2	2 / 9	1 / 8	4
8	3	8	1 / 5	5	4	2	9	3	1	2 / 3

TABLE 12

b_j / a_i	5	8	8	3	4	9	2	9	8	3
51	4	5	7	1	6	7	3	2	1	4
		8	8	3	1	9	2	9	8	3
8	0	5	8	2	3	5	9	4	2	6
	5				3					

TABLE 13

b_j / a_i	5	8	8	3	4	9	2	9	8	3
51	4	5	7	1	6	7	3	2	1	4
	5	8	8	3	4	1	2	9	8	3
8	2	0	2	1	5	2	2	4	3	2
						8				

As previously explained, the zero variables in rows 2 of A_2', A_3', A_4' and A_5' will remain zero in rows 2, 3, 4, and 5, respectively of the original problem A, independent of the (unknown optimal) values of a_i. The information presently available is displayed in Table 14.

TABLE 14

										51
x	x	x	x	x		x	x	x		a_2
x	x		x	x	x	x	x	x		a_3
	x	x	x		x	x	x	x	x	a_4
x	x	x	x	x		x	x	x	x	a_5
5	8	8	3	4	9	2	9	8	3	a_i / b_j

x represnts a variable whose value at optimality is known to equal zero.

Columns 2, 4, 7, 8, 9 may be removed. Row 5 may also be removed. The missing variable in row 5 will equal a_5, since all the other $x_{5j} = 0$. The reduced 4×5 problem appears in Table 15. In addition to the zeroes indicated by Table 14, the optimal values of the following variables are now known: $x_{12} = 8$; $x_{14} = 3$; $x_{17} = 2$, $x_{18} = 9$, $x_{19} = 8$, $x_{56} = a_5$.

TABLE 15

1	3	5	6	10	
4	7	6	7	4	21
2 x	5 x	8 x	1	1	a_2
3 x	1	4 x	2 x	2	a_3
10	8 x	3	5 x	6 x	a_4
5	8	4	$9 - a_5$	3	a_i / b_j

The exact values of the a_i will now be determined by trial and error. What has been gained is that instead of solving a 5×10 tableau many times, one solves a 4×5 tableau many times. The additional preliminary work of solving 4 long (2×10) transportation problems is not overly burdensome, and in addition to enabling a reduction in dimensionality, it provides knowledge regarding the exclusion of certain variables in the reduced tableau from the final solution.

BIBLIOGRAPHY

[1] Cooper, L., "Location-Allocation Problems," Operations Research *11*, 331-343 (1963).
[2] Cooper, L., "The Transportation-Location Problem," Operations Research *20*, 94-108 (1972).
[3] Dantzig, G.B., *Linear Programming and Extensions*, Princeton University Press, Princeton, N.J. (1963).
[4] Harris, B., "A Code for the Transportation Problem of Linear Programming," Journal of the Association for Computing Machinery, *23*, 155-157 (1976).
[5] Intrator, J. and J. Paroush, "Sensitivity Analysis of the Classical Transportation Problem," Computers and Operations Research *4*, 213-226 (1977).
[6] Kraemer, S., "Warehouse Location and Allocation Problems Solved by Mathematical Programming Methods," in *Operational Research 72*, edited by Miceal Ross, North Holland Publishing Co., Amsterdam (1973) pp. 541-549.
[7] Srinivasan, V. and G.L. Thompson, "An Operator Theory of Parametric Programming for the Transportation Problem—I and II," Naval Research Logistics Quarterly, *19*, 205-252 (1972).

ON A SEARCH FOR A MOVING TARGET

A. R. Washburn

Naval Postgraduate School
Monterey, California

ABSTRACT

We consider the problem of searching for a target that moves in discrete time and space according to some Markovian process. At each time, a searcher attempts to detect the target. If the searcher's action at each time is such as to maximize his chances of immediate detection, we call his strategy "myopic." We provide a computationally useful necessary condition for optimality, and use it to provide an example wherein the myopic strategy is not optimal.

DESCRIPTION OF THE PROBLEM

Let X be the position space within which the target moves, and let $P(x,1)$ be the probability that the target starts at position x at time 1 (so $\sum_{x \in X} P(x,1) = 1$). In general, let $P(x,t)$ be the probability that the target is at position x at time t and has not been detected by any of the searches at times $1, 2, \ldots, t-1$; $x \in X$, $1 \leqslant t \leqslant m$. If $q(x,t)$ is the probability that a target at position x at time t will not be detected at time t, then $\tilde{P}(x,t) = P(x,t)\, q(x,t)$ is the probability that the target is at x at t and has not been detected by any of the searches at times $1, \ldots, t$. If $q(x,t)$ is selected to minimize $\sum_{x \in X} \tilde{P}(x,t)$ within whatever constraints are imposed on the search, then the search is *myopic*. After the search at time t, the target moves to its position at time $t+1$; thus

$$P(y,t+1) = \sum_{x \in X} \tilde{P}(x,t)\, \Gamma(x,y,t);\ y \in X,\ 1 \leqslant t \leqslant m-1, \tag{1}$$

where $\Gamma(x,y,t)$ is the probability that the target is at y at $t+1$, given that it is at x at t.

The problem is to select $q(x,t)$, $x \in X$ and $1 \leqslant t \leqslant m$, so as to minimize $s_m = \sum_{x \in X} \tilde{P}(x,m)$, where s_m is just the probability that the target is never detected at any time $\leqslant m$. We will introduce a necessary condition for optimality in problems of this sort that lends itself to successive improvement of search strategies, and will give an example showing that the myopic strategy can be improved.

BACKGROUND

Pollock [9] considered a special case of this problem where there were two positions and where the searcher had to decide which one to search at each time. He solved it using an application of dynamic programming that is not easily extended. Dobbie [4] subsequently solved a

similar problem where time was continuous, rather than discrete. Iida [8] obtained necessary and sufficient conditions for optimality when the target's motion is discrete but not necessarily Markov, when the detection function is regular [12], and when effort is infinitely divisible. He used his condition to solve a problem where there were five essentially different paths that the target might follow. Hellman [7] and Saretsalo [11] have also derived necessary conditions for optimality in continuous diffusion and in general Markov motion, respectively. All of these conditions differ from ours, and are apparently not as amenable to computation.

The author became interested in this subject after reading Hellman [5,7], who showed that the effect of random search on a target whose motion is a diffusion is to introduce an extra term into the diffusion equation for the probability density. After making a discrete approximation to the diffusion equation, the author based an algorithm for finding the optimal effort distribution on a necessary condition involving the adjoint of the discrete diffusion equation, and submitted a paper on the technique to this journal in 1976.

During a revision of this paper, Brown [1] appeared and gave necessary and sufficient conditions for optimality in the case of arbitrary target motion in discrete time and space but with an exponential detection function. Brown derived an algorithm for optimizing search plans which, like his conditions, was based on planning search at a given time interval using a target location distribution which accounted for failure to detect the target both with past and with future search. This algorithm is based on the simple but powerful condition involving the function that is $Q(x,t,q)$ in this paper, representing the effect of future search, and playing the same algorithmic role as the solution of the adjoint equation in the author's earlier submission. In the present paper, we observe that a slight generalization of Brown's condition for Markov processes is necessary for optimality when the detection function is arbitrary, as long as glimpses at different times are independent. In particular, we permit the restriction that all effort must be placed in a single cell at each opportunity, which is sometimes a restriction in practice.

THE NECESSARY CONDITION

To emphasize the dependence of $P(x,t)$ on the search strategy $q(\cdot,\cdot)$, we will write $P(x,t,q)$; however, $P(\cdot,t,q,)$ does not depend on $q(\cdot,\tau)$ for $\tau \geqslant t$. Define $Q(x,t,q)$ to be the probability that the target is not detected by any of the searches at $t+1, \ldots, m$, given that the target is at x at time t, with $Q(\cdot,m,\cdot) \equiv 1$. $Q(\cdot,t,q)$ does not depend on $q(\cdot,\tau)$ for $\tau \leqslant t$.

For any t between 1 and m, the probability that the target is not detected by any of the first m searches is, according to the theorem of total probability,

$$s_m = \sum_{x \in X} P(x,t,q)\ q(x,t)\ Q(x,t,q). \tag{2}$$

From our earlier observations, the product $P(x,t,q)\ Q(x,t,q)$ does not depend on $q(\cdot,t)$. Let $S(q,t)$ be the set of functions $f(\cdot)$ that are feasible for $q(\cdot,t)$ when $q(\cdot,\tau)$ is specified for $\tau \neq t$. Then, if s_m is minimal, it must be true that $\sum_{x \in X} P(x,t,q)\ f(x)\ Q(x,t,q)$ is minimized for $f \in S(q,t)$ when $f(\cdot) = q(\cdot,t)$. This is our necessary condition. The computational usefulness of the condition arises from the fact that functions of position only, rather than position and time, are involved in the optimization. Strategies satisfying the necessary condition will be referred to as "critical."

STRATEGY IMPROVEMENT

Given any search strategy $q(\cdot,\cdot)$ that does not satisfy the necessary condition for some t^*, it is evident that a better strategy is $q'(\cdot,t) = q(\cdot,t)$ for $t \neq t^*$, and $q'(\cdot,t^*) = f(\cdot)$, where $f(\cdot)$ is the minimizing function in $S(q,t^*)$. Repetitive applications will always result in (strictly) improved strategies as long as the necessary condition is not satisfied. The flow diagram in Fig. 1 simply organizes this procedure in a manner that is computationally efficient. It assumes that $Q(\cdot,\cdot,q)$ is initially known. The efficient thing about the procedure is that, even though there are potentially m distinct search strategies considered, only m applications of Eq. (1) and no computations on Q are needed. The reason for this is that time is considered sequentially, so that in the minimization step $P(x,t,q)$ depends only on the part of $q(\cdot,\cdot)$ that has been changed, whereas $Q(x,t,q)$ depends only on the part of $q(\cdot,\cdot)$ that has not been changed. Note that the search strategy achieved at EXIT would be the myopic strategy if $Q(\cdot,\cdot,q)$ were set identically equal to 1 at the input.

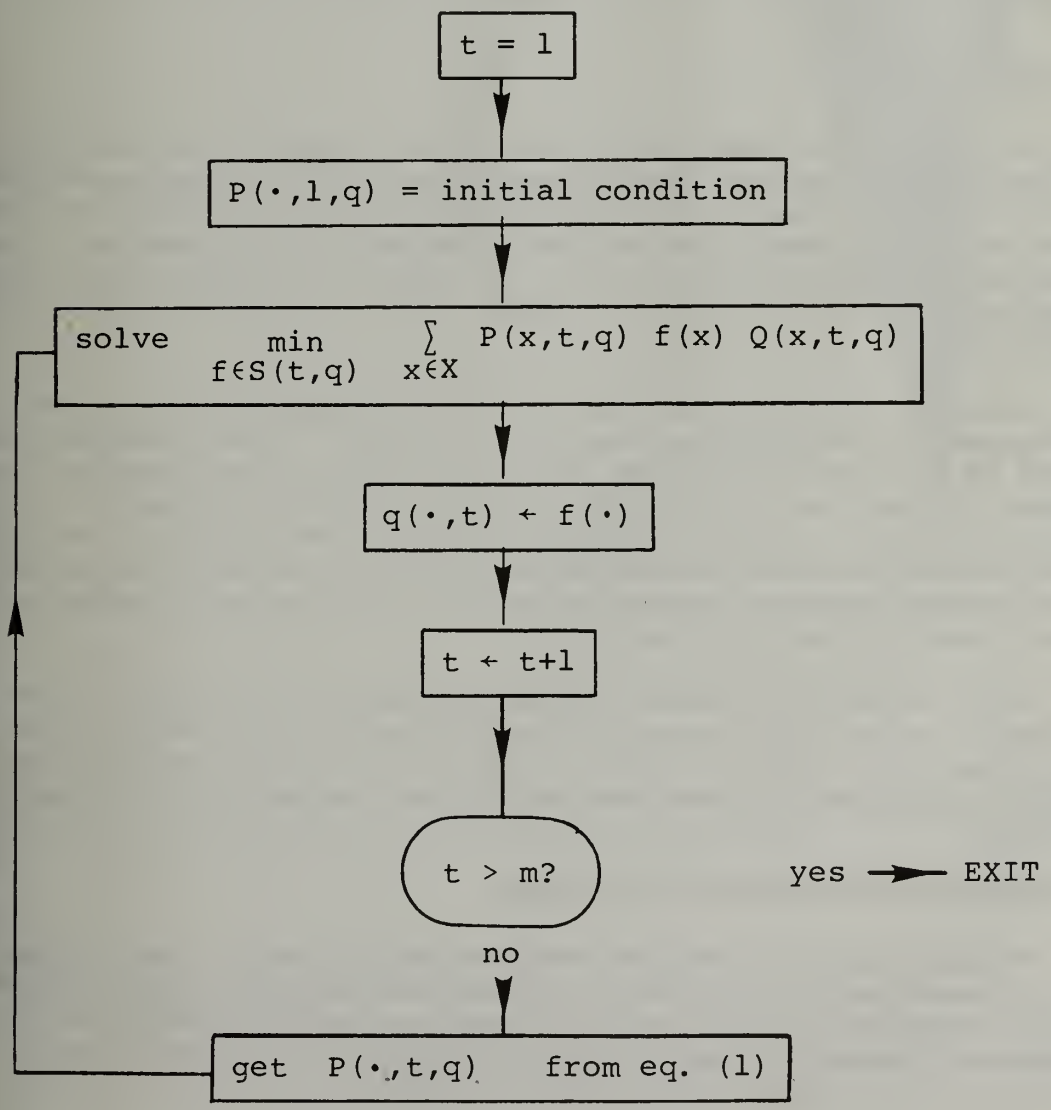

FIGURE 1. Flow diagram for strategy improvement.

Repeated passes through the flow diagram will result in a succession of strategies, each strictly better than the preceding as long as the preceding does not satisfy the necessary condition. Before each pass, the function $Q(\cdot,\cdot,q)$ must be computed. This can be done recursively by initializing $Q(\cdot,m,q) = 1$, $x \in X$, and then making $m - 1$ applications of Eq. (3):

$$Q(x,t-1,q) = \sum_{y \in X} \Gamma(x,y,t-1)\ q(y,t)\ Q(y,t,q), \tag{3}$$

$$x \in X,\ t = m,\ m-1,\ \ldots,\ 2.$$

MYOPIC STARTING POINT

The present writer and also Brown [1] have found that one has to be ingenious to think of examples where the myopic strategy differs significantly in terms of detection probability from strategies that satisfy the necessary condition. The problem in the next section will provide a typical example where the difference is small. Based on this experience, it is reasonable to conclude that the myopic strategy provides a good starting point for the iterative procedure discussed in the previous section. This can be done by beginning the first iteration with $Q(\cdot,\cdot,q) = 1$.

AN EXAMPLE

Let X be the first 67 integers with the target initially on number 34. The target moves in a discrete diffusion; from interior points it goes left or does not move or goes right with probabilities 0.3, 0.4, 0.3, respectively. From either of the two end positions, the target stays where it is or returns to the closest interior point with probabilities 0.4 or 0.6, respectively. At each opportunity, the searcher picks any one of the 67 integers; if the target is currently at that integer, the probability that it will not be detected is 0.875, or otherwise 1.0. There are 80 opportunities.

If the searcher were to pick an integer at random at each opportunity, the probability of nondetection for all 80 opportunities would be $(1 - 0.125/67)^{80} = 0.8612$. Successive passes through the flow diagram in Fig. 1, with $Q(\cdot,\cdot,q) \equiv 1$ initially, produce probabilities of nondetection of 0.3664, 0.3641, 0.3629, 0.3626, and 0.3625, with 0.3664 corresponding to myopic search (q_1) and with the strategy corresponding to 0.3625 (q_2) being critical. q_2 is not globally optimal in this problem; after trying some different starting points, another strategy (q_3) satisfying the necessary condition with nondetection probability 0.3623 was found. A globally optimal search strategy for this problem is unknown.

Figure 2 shows q_1 and contrasts it with q_2. q_2 is somewhat more "spread out" than q_1 because $Q(\cdot,\cdot,q_2)$ is smallest for values of x near the center. $P(\cdot,\cdot,q_2)$ and $Q(\cdot,\cdot,q_2)$ can be seen in Figs. 3 and 4. Figure 5 contrasts $P(x,40,q_1)$ with $P(x,40,q_2)$. The myopic strategy tends to cause a flat spot in the region being searched that is missing when critical strategies are employed.

EXTENSIONS AND POSSIBLE EXTENSIONS

Permitting space to be continuous rather than discrete is simply a matter of replacing sums by integrals and $\Gamma(x,y,t)$ by $\Gamma(x,dy,t)$ in the preceding. Permitting time to be continuous would be more difficult, since one would have to begin by changing the definition of $q(x,t)$ and abandoning the simple step-by-step proofs that are possible when time is discrete. We will not speculate on the form of the necessary condition in this case.

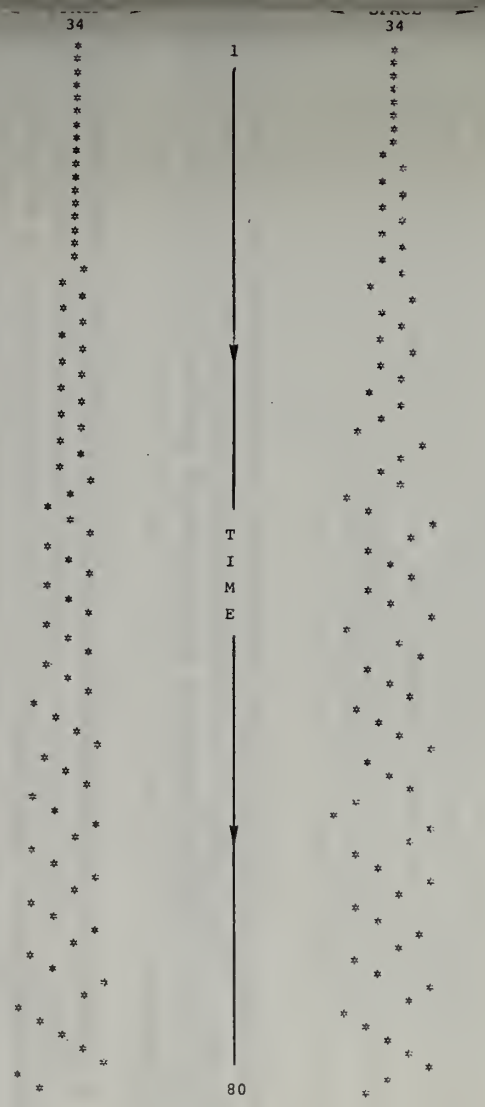

FIGURE 2. Comparison of myopic (left) and critical (right) strategies.

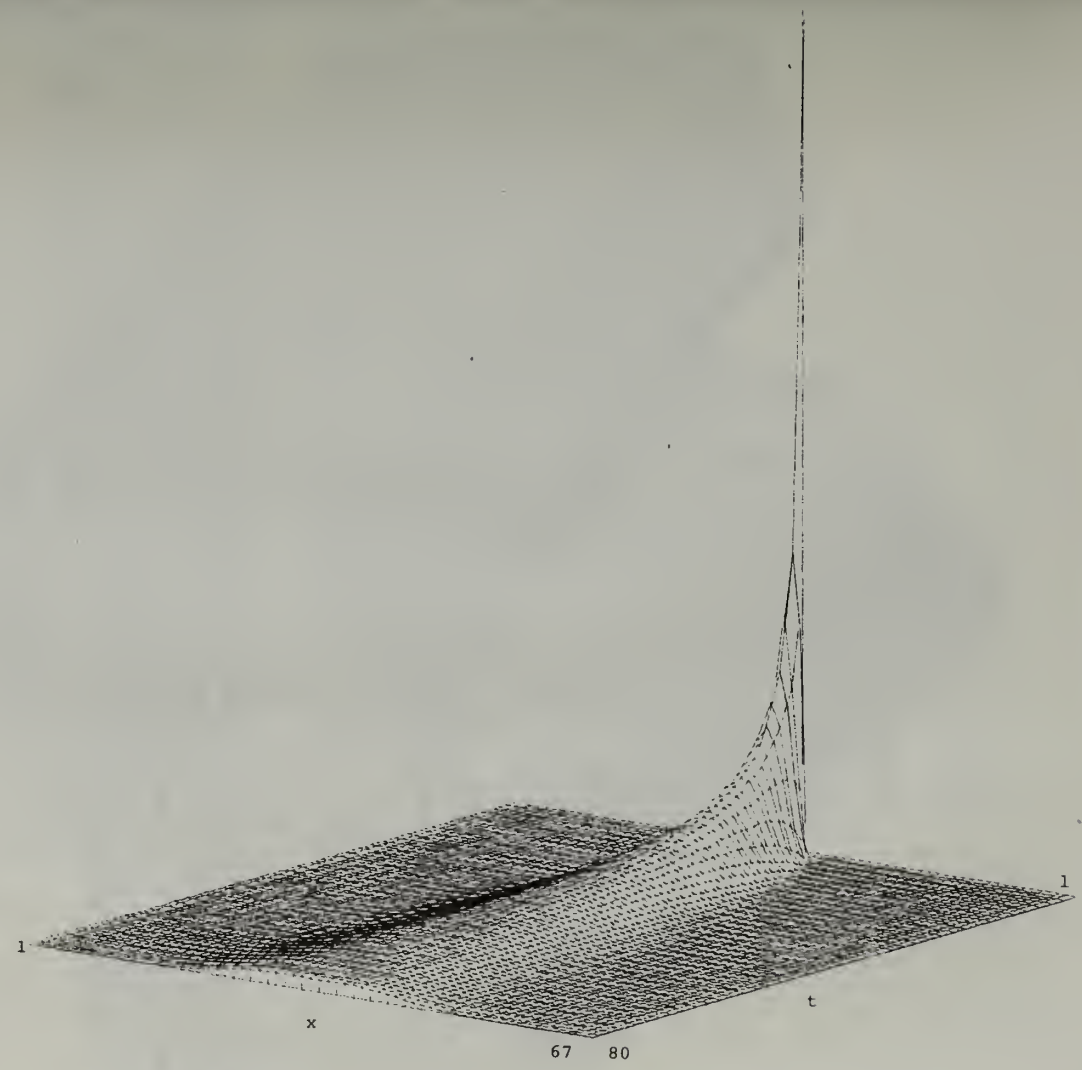

FIGURE 3. The function $P(x,t,q_2)$.

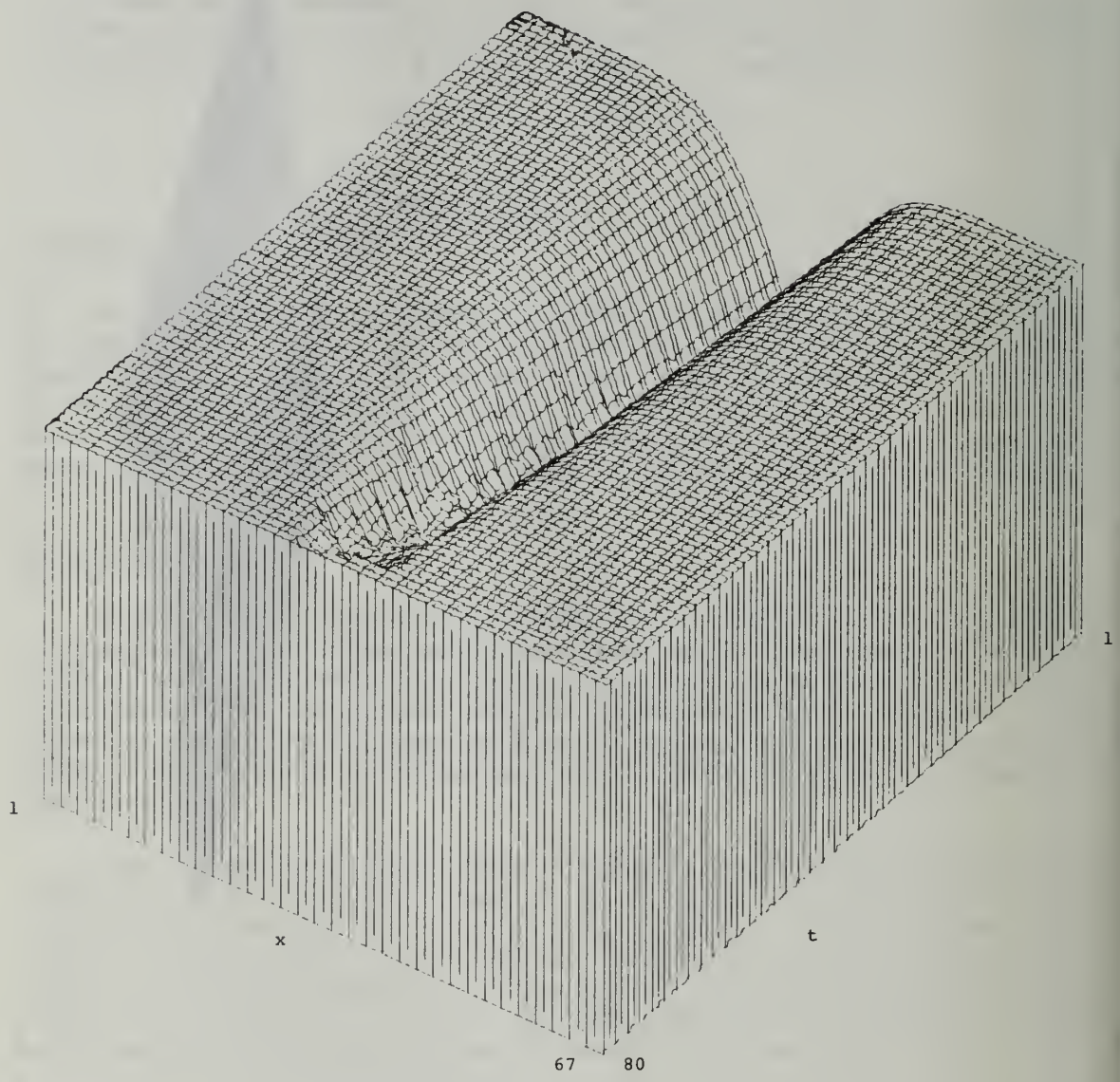

FIGURE 4. The function $Q(x,t,q_2)$.

The necessary condition would be more useful if it were also sufficient. Brown [1] has shown that it is in fact sufficient when the detection law [12] is exponential and a fixed amount of search effort is available at each time. His proof would hold for any strictly convex, decreasing detection law.

In search for a stationary target, the myopic strategy is optimal for both maximizing detection probability at a fixed time and for minimizing the expected time to detection [12]. The myopic strategy can be implemented by providing a display of $P(\cdot,t,q)$ to an operator who presumably selects positions x at which $P(x,t,q)$ is large at which to continue the search; computer-aided search schemes such as [2] function in this manner even when the search is for a moving target. Strictly speaking, one ought to be explicit about the objective function when searching for a moving target, since myopic search is not generally optimal for the probability of detection in a fixed time problem when the target moves. Whether myopic search is optimal

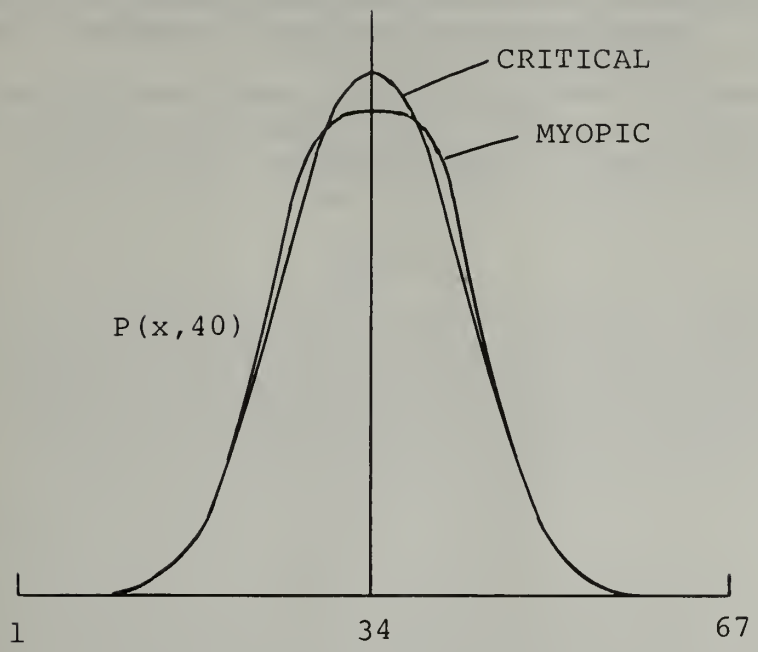

FIGURE 5. Comparison of $P(x,40,q_1)$ with $P(x,40,q_2)$.

or even near optimal for the expected-time-to-detection problem is unknown. This is unfortunate, since expected-time-to-detection is a reasonable criterion in problems where the target is very valuable and search cost is proportional to time spent searching. At the present time, to the best of this author's knowledge, there is no algorithm applicable to the problem of computing the search strategy that provides the minimal expected time-to-detection. Construction of such an algorithm would be a useful extension.

BIBLIOGRAPHY

[1] Brown, S., "Optimal and Near Optimal Search for a Target with Multiple Scenario Markovian, Constrained Markovian, or Geometric Memory Motion in Discrete Time and Space," Daniel H. Wagner Associates, Memorandum Report, June 14, 1977.

[2] Dicenza, J. and H. Richardson, "Operational Use of the Computer in Coast Guard Search Planning," presented at the joint ORSA/TIMS meeting in Las Vegas (1975).

[3] Dobbie, J., "A Survey of Search Theory," Journal of the Operations Research Society of America *16*, 525-537 (1968).

[4] Dobbie, J., "A Two-Cell Model of Search for a Moving Target," Journal of the Operations Research Society of America *22*, 79-92 (1974).

[5] Hellman, O.B., "On the Effect of Search Upon the Probability Distribution of a Target Whose Motion is a Diffusion Process," Annals of Mathematical Statistics *41*, 1717-1724 (1970).

[6] Hellman, O.B., "Optimal Search for a Randomly Moving Object in a Special Case," Journal of Applied Probability *8*, 606-661 (1971).

[7] Hellman, O.B., "On the Optimal Search for a Randomly Moving Target," SIAM Journal of Applied Mathematics *22*, 545-552 (1972).

[8] Iida, Koji, "Ido Mokuhyobutsu no Tansaku (The Optimal Distribution of Searching Effort for a Moving Target)," Keiei Kagaku (Japan) *16*, 204-215 (1972), in Japanese.

[9] Pollock, S.M., "A Simple Model of Search for a Moving Target," Journal of the Operations Research Society of America *18*, 883-903 (1970).

[10] Pursiheimo, U., "On the Optimal Search for a Moving Target," Report 35, Institute for Applied Mathematics, University of Turku, Finland (1972).
[11] Saretsalo, L., "On the Optimal Search for a Target Whose Motion is a Markov Process," Journal of Applied Probability *10*, 847-856 (1973).
[12] Stone, Lawrence D., *Theory of Optimal Search*, Academic Press, New York (1975).
[13] Stone, L. and H. Richardson, "Search for Targets with Conditionally Deterministic Motion," SIAM Journal of Applied Mathematics *27*, 239-255 (1974).

ANALYSIS OF DATA FROM LIFE-TEST EXPERIMENTS UNDER AN EXPONENTIAL MODEL

J. F. Lawless and K. Singhal

University of Waterloo
Ontario, Canada

ABSTRACT

This paper discusses situations in which the distribution of a lifetime response variable T is taken to depend upon a vector $\underline{x}$ of regressor variables. We specifically consider the case in which T, given $\underline{x}$, has an exponential distribution, and in which $\underline{x}$ represents levels of fixed factors in an experimental design. Methods of analyzing data under this type of model are discussed, with maximum likelihood and least squares methods being presented and compared.

1. INTRODUCTION

In many life-testing situations the life distribution of the items under study is dependent upon physical and environmental factors. For example, the lifetime of a capacitor may depend upon the voltage and temperature the capacitor is subject to, the time to breakdown of a type of electrical insulation may depend on voltage as well as certain physical characteristics of the insulation, and so on. With such items, life-test experiments are often carried out in which several factors affecting lifetime are varied simultaneously. This paper deals with the analysis of data from such experiments. We deal specifically with situations in which the life distribution of the items under consideration is, at fixed environmental conditions, exponential. We also assume that the scale parameter in the exponential distribution is related to the environmental factors in a multiplicative fashion. That is, the lifetime T of an item depends on a given set of environmental factors which can be represented by a (row) vector $\underline{x}$. The distribution of T, given $\underline{x}$, is exponential with density

$$\theta_{\underline{x}}^{-1} \exp(-t/\theta_{\underline{x}}), \qquad t > 0, \tag{1}$$

where $\theta_{\underline{x}} = \exp(\underline{x}\underline{\beta}) = \exp(x_1\beta_1 + \ldots + x_k\beta_k)$. This paper deals mainly with analysis of variance type models, in which θ depends in a multiplicative way on fixed factors, in which case it is often convenient to write the model in a form (see Eq. (4)) analogous to that used in normal theory analysis of variance.

The model described above covers a wide variety of applications. Before proceeding we briefly note some examples.

EXAMPLE 1: Zelen [16,17] has considered factorial experiments with an exponential model; for a two-way model, for example, we have the mean life θ depending on two factors A and B, having a and b levels, respectively. His assumed model has

$$\theta_{ij} = m a_i b_j c_{ij}, \; i = 1, \ldots, a \; ; \; j = 1, \ldots, b, \tag{2}$$

where the parameters are subject to restrictions

$$\prod_{i=1}^{a} a_i = \prod_{j=1}^{b} b_j = \prod_{i=1}^{a} c_{ij} = \prod_{j=1}^{b} c_{ij} = 1. \tag{3}$$

This model can be written in the form (1), with

$$\theta_{ij} = \exp(\mu + \alpha_i + \beta_j + \gamma_{ij}), \tag{4}$$

where $\mu = \log m$, $\alpha_i = \log a_i$, $\beta_j = \log b_j$, and $\gamma_{ij} = \log c_{ij}$. In this paper we generalize and extend Zelen's method of analysis for this model.

EXAMPLE 2: Power law models play a large role in many engineering studies. For example, Nelson [12] and Lawless [9] have considered data on time to breakdown of electrical insulating fluid under constant voltage v. The assumed model involves a power rule, with $\theta = cv^{-P}$; letting $x = \log v$, this can be expressed in the form (1), with $\theta = \exp(\alpha + \beta x)$, where $\alpha = \log c$ and $\beta = -p$.

EXAMPLE 3: All four models discussed on pp. 421-422 of Mann et al. [11] can be expressed in the form (1). For example, the Generalized Eyring Model [11, p. 422] has $\theta^{-1} = AT\,[\exp(-B/kT)]\,[\exp(CV + DV/kT)]$, where T and V are environmental stresses, A, B, C, D are unknown parameters, and k is Boltzmann's constant. This model can be written in the form (1) with

$$\theta = \exp[-\log A - \log T + B(1/kT) - CV - D(V/kT)].$$

Several authors have considered the analysis of data from models of the form (1). For example, Lawless [9], Prentice [13], Singpurwalla [14], and Kahn [5] discuss the particular form of Eq. (1) mentioned in Example 2 previously, while Lawless and Singhal [10], Prentice [13], and Singpurwalla et al. [15] discuss models (1) with more than one regressor variable present. This paper deals primarily with analysis of variance type models such as Eq. (4): these can, of course, be treated as special cases of Eq. (1) and analyzed in this way, though it is sometimes useful to dwell on the particular "fixed-effects" form of Eq. (4). This has received less attention in the literature, with the main results to this point given by Zelen [16,17]. In this paper we extend Zelen's methods of analysis and discuss alternate methods. In particular, Zelen has discussed the model (4) in the special case in which the experiment is run so that at each combination of factor levels (i,j) only the first r failure times are observed. He shows how to carry out likelihood ratio tests concerning various factors in the model, and how to obtain confidence intervals for ratios of different a_i's and b_j's in Eq. (2). We extend his work by considering the case where the first r_{ij} failure times are observed at levels (i,j). We also consider more detailed analyses of data from the model (4), involving the examination of treatment contrasts, for example.

2. TWO-FACTOR EXPERIMENTS

The remainder of the paper will be mainly concerned with analyzing data from a two-factor model of the form (4). This allows us to discuss all relevant features of proposed methods of analysis; the techniques are, however, easily used with other models of the form (1). The two-factor model has previously been discussed by Zelen [16,17], who gives methods of analysis based on likelihood ratio tests. We extend (and correct, in one instance) his work here, and also discuss alternate methods of analysis.

We consider experiments in which there are two factors A and B, having a and b levels, respectively, with experimentation at all ab factor level combinations. At the levels (i,j) for A

and B, n_{ij} items are simultaneously put on test, the experiment being terminated at the time of the r_{ij}th item failure. Let t_{ijk} be the k th failure time for combination (i,j) $(k = 1,\ldots,r_{ij})$; the t_{ijk}'s are the first r_{ij} ordered observations in a sample of size n_{ij} from an exponential distribution with mean θ_{ij} $(i = 1,\ldots,a\,;j = 1,\ldots,b)$. We consider the model (4) where, in the most general case,

$$\log \theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \tag{5}$$

There are only ab independent parameters in Eq. (5), and for convenience we take the linear restrictions on the parameters to be

$$\sum_{i=1}^{a} r_{i.}\,\alpha_i = \sum_{j=1}^{b} r_{.j}\beta_j = 0 \tag{6}$$

$$\sum_{i=1}^{a} r_{ij}\gamma_{ij} = 0 \text{ (for all } j), \quad \sum_{j=1}^{b} r_{ij}\gamma_{ij} = 0 \text{ (for all } i).$$

Here $r_{i.} = \sum_{j=1}^{b} r_{ij}$, $r_{.j} = \sum_{i=1}^{a} r_{ij}$; we will later also use $r.. = \sum_{i=1}^{a}\sum_{j=1}^{b} r_{ij}$.

It is well known [4] that the quantities

$$T_{ij} = \sum_{k=1}^{r_{ij}} t_{ijk} + (n_{ij} - r_{ij})\,t_{ijr_{ij}}$$

are sufficient for the θ_{ij}'s, and that T_{ij} has a gamma distribution with density

$$f(T_{ij};\theta_{ij}) = \frac{T_{ij}^{\,r_{ij}-1}}{(r_{ij}-1)!\,\theta_{ij}^{\,r_{ij}}} \exp(-T_{ij}/\theta_{ij}).$$

The likelihood function for the model is

$$L(\theta_{ij}\text{'s}) = \prod_{i=1}^{a}\prod_{j=1}^{b} f(T_{ij};\,\theta_{ij})$$

and the log likelihood function is, except for an additive constant which we omit without loss of generality,

$$\log L = -\sum_{i=1}^{a}\sum_{j=1}^{b} r_{ij}\log\theta_{ij} - \sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij}/\theta_{ij}. \tag{7}$$

We now discuss methods of analyzing data from this model. We begin by considering likelihood ratio tests of various factors in the model, and then proceed to a discussion of maximum likelihood and least squares procedures.

3. LIKELIHOOD RATIO TESTS OF EFFECTS

In analyzing data from the model in Section 2, a number of hypotheses will be of interest. These include hypotheses of "no interaction" between factors, and hypotheses concerning "main effects." To formulate these ideas, we consider a set of five hypothesized models:

$$H_4 : \log\theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

$$H_3 : \log\theta_{ij} = \mu + \alpha_i + \beta_j,$$

$$H_2 : \log\theta_{ij} = \mu + \beta_j,$$

$$H_1 : \log \theta_{ij} = \mu + \alpha_i,$$
$$H_0 : \log \theta_{ij} = \mu.$$

In these models, the α_i's, β_j's, and γ_{ij}'s satisfy the linear restrictions given in Eq. (6). Preliminary analysis of data will often involve testing some of these models against others. For example, to test for "no interaction" between A and B, we test H_3 vs H_4, taking H_3 as the null and H_4 as the alternative hypothesis. In general, to test H_s (null hypothesis) vs H_t (alternative), we can use the generalized likelihood ratio test statistic

$$\lambda_{st} = -2 \log [L_{\max}(H_s)/L_{\max}(H_t)]$$
$$= 2(\Lambda_s - \Lambda_t),$$

where $\Lambda_s = -\log L_{\max}(H_s)$ is minus the maximum of $\log L$ under the model H_s. Under H_s, λ_{st} is asymptotically (i.e., as the r_{ij}'s become large) χ^2 with degrees of freedom given by the difference in the number of functionally independent parameters in H_t and H_s (e.g. [7] Ch. 24).

We now give the maximum likelihood estimates (m.l.e.'s) and maximized log likelihoods necessary for carrying out likelihood ratio tests for models H_0 to H_4.

H_4: Under H_4, there is a one-to-one relationship between $\{\theta_{ij}\}$ and $\{\mu, \alpha_i, \beta_j, \gamma_{ij}\}$, determined by Eqs. (5) and (6). It is well known [4] that the maximum likelihood estimate of θ_{ij} is $\hat{\theta}_{ij} = T_{ij}/r_{ij}$. We find that (henceforth all sums over i are from 1 to a and over j are from 1 to b, unless stated otherwise) the parameter estimates $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are given by the following equations:

i) $$\hat{\mu} = \frac{1}{r\,..} \sum_i \sum_j r_{ij} \log \hat{\theta}_{ij},$$

ii) $\hat{\gamma}_{ij} = \log \hat{\theta}_{ij} - \hat{\alpha}_i - \hat{\beta}_j - \hat{\mu}$, where

iii) The $\hat{\alpha}_i$'s and $\hat{\beta}_j$'s are found as solutions to the linear equations

(9)
$$\sum_j r_{ij} \log \hat{\theta}_{ij} = r_{i.}(\hat{\mu} + \alpha_i) + \sum_j r_{ij}\beta_j, \qquad i = 1, \ldots, a,$$

$$\sum_i r_{ij} \log \hat{\theta}_{ij} = r_{.j}(\hat{\mu} + \beta_j) + \sum_i r_{ij}\alpha_i, \qquad j = 1, \ldots, b,$$

subject to the restrictions $\sum_i r_{i.}\alpha_i = \sum r_{.j}\beta_j = 0$. Comments on the solutions of such equations are given, for example, by Kempthorne [6, pp. 80-81]. From the m.l.e.'s we obtain

$$\Lambda_4 = \sum_i \sum_j r_{ij} \log \hat{\theta}_{ij} + \sum_i \sum_j T_{ij}/\hat{\theta}_{ij}$$
$$= r\,..\hat{\mu} + r\,..$$

Note, in particular, that in order to obtain Λ_4, it is not necessary to obtain anything other than $\hat{\mu}$. This is noteworthy, since often (see following) we may wish to test H_3 against H_4, and then to discuss only H_3 further, if H_3 is accepted.

H_3: The maximum likelihood estimates of μ, α_i, and β_j under H_3 cannot be given in closed form, and must be determined numerically. To maximize $\log L$ under the restriction on

the α_i's and β_j's we consider

$$\log L = -r\,.\,.\,\mu - \sum_i \sum_j T_{ij} \exp(-\mu - \alpha_i - \beta_j),$$

where we consider, through Eq. (6), that $\alpha_a = -\frac{1}{r_{a.}} \sum_{i \neq a} r_{i.}\,\alpha_i$ and $\beta_b = -\frac{1}{r_{.b}} \sum_{j \neq b} r_{.j}\beta_j$. That is, $\mu, \alpha_1, \ldots, \alpha_{a-1}$ and $\beta_1, \ldots, \beta_{b-1}$ are considered as the unknown parameters and α_a and β_b are given in terms of these. Differentiation of $\log L$ gives equations

$$\frac{\partial \log L}{\partial \mu} = -r\,.\,. + \sum_i \sum_j T_{ij} \exp(-\mu - \alpha_i - \beta_j) = 0,$$

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_j T_{ij} \exp(-\mu - \alpha_i - \beta_j) - \frac{r_{i.}}{r_{a.}} \sum_j T_{aj} \exp(-\mu - \alpha_a - \beta_j), \tag{10}$$

$$i = 1, \ldots, a-1,$$

$$\frac{\partial \log L}{\partial \beta_j} = \sum_i T_{ij} \exp(-\mu - \alpha_i - \beta_j) - \frac{r_{.j}}{r_{.b}} \sum_i T_{ib} \exp(-\mu - \alpha_i - \beta_b),$$

$$j = 1, \ldots, b-1.$$

These equations have to be solved iteratively. This is easily done, for example, using Newton's method [7, Ch. 18]. We note that again in this case, Λ_3 is of the form

$$\Lambda_3 = r\,.\,.\hat{\mu} + r\,.\,.,$$

where $\hat{\mu}$ is the maximum likelihood estimate of μ obtained from expression (10). To see this, observe that the first equation in expression (10) implies $\sum_i \sum_j T_{ij} \exp(-\hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = r\,.\,.$, so that substituting the m.l.e.'s in $\log L$ gives Λ_3 as stated.

H_2: In this case we need to consider

$$\log L = -r\,.\,.\mu - \sum_i \sum_j T_{ij} \exp(-\mu - \beta_j)$$

where again, using Eq. (6), we treat $\beta_1, \ldots, \beta_{b-1}$ as the unknown parameters, with $\beta_b = -\frac{1}{r_{.b}} \sum_{j \neq b} r_{.j}\beta_j$. It is easily found that $\log L$ is maximized for μ and $\beta_1, \ldots, \beta_b$ by choosing

$$\hat{\mu} = \frac{1}{r\,.\,.} \sum_j r_{.j} \log(T_{.j}/r_{.j})$$

and

$$\hat{\beta}_j = \log(T_{.j}/r_{.j}) - \hat{\mu}, \qquad j = 1, \ldots, b,$$

where $T_{.j} = \sum_i T_{ij}$. Λ_2 then once again equals $r\,.\,.\hat{\mu} + r\,.\,.$.

H_1: We similarly find here that maximizing $\log L$ leads to estimates

$$\hat{\mu} = \frac{1}{r\,.\,.} \sum_i r_{i.} \log(T_{i.}/r_{i.})$$

$$\hat{\alpha}_i = \log(T_{i.}/r_{i.}) - \hat{\mu}, \qquad i = 1, \ldots, a,$$

where $T_{i.} = \sum_j T_{ij}$. Then, $\Lambda_1 = r\,.\,.\hat{\mu} + r\,.\,.$.

H_0: In this case we find $\hat{\mu} = \log(T.\,./r.\,.)$ and $\Lambda_0 = r..\hat{\mu} + r..$, where $T.\,. = \sum_i \sum_j T_{ij}$.

Utilizing the expressions given above, we can test any form of the model against any other. We will often want to start by testing H_3 vs H_4. Acceptance of H_3 would mean that there is no interaction, in a sense, between factors A and B, and this would make it meaningful to consider questions about the α_i's and β_j's individually. If H_3 is accepted we might, for example, then want to examine the hypothesis that Factor A has no effect. This would be done by testing H_2 vs H_3. Similarly, we might want to test H_1 vs H_3. The likelihood ratio method is a convenient way to do this, and examples of the above tests are given in Section 6.

REMARK: Zelen [17] gives likelihood ratio tests like those discussed here, for the case in which all r_{ij}'s are equal. It should be noted that the statistic M_3 given on p. 513 of his paper is incorrect, however, as is the expression for $\tilde{m}_3$ on p. 512. This error arises because Zelen obtains incorrect expressions for his parameter estimates under model H_3. In fact, he uses closed-form expressions for the estimates, though as noted above, it is not possible to obtain these.

4. LARGE SAMPLE MAXIMUM LIKELIHOOD PROCEDURES

Likelihood ratio procedures are convenient for testing hypotheses concerning main effects and interactions in the model (5), but are less convenient when, for example, it is desired to estimate contrasts involving the parameters or to obtain confidence intervals for parameters. We discuss here procedures based on large sample properties of the m.l.e.'s, which can be used to test overall effects in the model or for estimation purposes.

We begin by considering the full model (5), in which case the log likelihood is, from Eq. (7),

$$\log L(\mu, \alpha_i, \beta_j, \gamma_{ij}) = -\sum_{i=1}^{a}\sum_{j=1}^{b} r_{ij}(\mu + \alpha_i + \beta_j + \gamma_{ij}) - \sum_{i=1}^{a}\sum_{j=1}^{b} T_{ij} \exp(-\mu - \alpha_i - \beta_j - \gamma_{ij}).$$

The parameters satisfy the linear restrictions (6). In what follows we will suppose that the ab unknown parameters in the model are $\mu, \alpha_1, \ldots, \alpha_{a-1}, \beta_1, \ldots, \beta_{b-1}$, and γ_{ij} $(i = 1, \ldots, a-1; j = 1, \ldots, b-1)$. The remaining parameters can be given in terms of these through Eq. (6).

Using standard asymptotic properties of m.l.e.'s, the covariance matrix of the limiting normal distribution for $(\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_{a-1}, \hat{\beta}_1, \ldots, \hat{\beta}_{b-1}, \hat{\gamma}_{11}, \ldots, \hat{\gamma}_{(a-1)(b-1)})$ is given by the inverse of the matrix of negative expected second deviatives of $\log L$. This is found here to be of the form

$$V = \begin{pmatrix} r.. & 0 & 0 & 0 \\ 0 & A & D & 0 \\ 0 & D' & B & 0 \\ 0 & 0 & 0 & C \end{pmatrix}^{-1}, \tag{11}$$

where the estimates are written in the order given above, and V is therefore partitioned so that A is $(a-1) \times (a-1)$, B is $(b-1) \times (b-1)$, and C is $(a-1)(b-1) \times (a-1)(b-1)$. Specifically, the entries in the matrices A, B, and D are

$$A_{ij} = E\left(\frac{-\partial^2 \log L}{\partial\alpha_i \partial\alpha_j}\right) = \frac{r_{i.} r_{j.}}{r_{a.}} \; (i \neq j); \; A_{ii} = r_{i.} + \frac{r_{i.}^2}{r_{a.}},$$

12) $$B_{ij} = E\left(\frac{-\partial^2 \log L}{\partial\beta_i \partial\beta_j}\right) = \frac{r_{.i}r_{.j}}{r_{.b}} \ (i \neq j); \ B_{jj} = r_{.j} + \frac{r_{.j}^2}{r_{.b}},$$

$$D_{ij.} = E\left(\frac{-\partial^2 \log L}{\partial\alpha_i \partial\beta_j}\right) = r_{ij} - \frac{r_{.j}r_{ib}}{r_{.b}} - \frac{r_{i.}r_{aj}}{r_{a.}} + \frac{r_{i.}r_{.j}r_{ab}}{r_{a.}r_{.b}}.$$

The expression for C_{ij} is more complicated and will not be given here, since we do not require an explicit expression for it in the sequel. The formula for C_{ij} is, however, precisely the same as the expression given in Kendall and Stuart [8, Section 35.79] for the case of least squares regression with unequal replication. We also note that in the "proportional frequency" case, when $r_{ij} = r_{i.}r_{.j}/r\ldots$, things simplify to a degree, since then $E(-\partial^2 \log L/\partial\alpha_i\partial\beta_j) = 0$, so that $D = 0$.

An alternate way to obtain expression (11) is by analogy with least squares methods for unbalanced two-way models. Since $\hat{\theta}_{ij} = T_{ij}/r_{ij}$ has a gamma distribution, $\log \hat{\theta}_{ij}$ has a log gamma distribution (e.g., see Bartlett and Kendall [2]) with mean and variance given by

(13) $$E(\log \hat{\theta}_{ij}) = \log \theta_{ij} - \log r_{ij} + \psi(r_{ij}),$$

$$Var(\log \hat{\theta}_{ij}) = \psi'(r_{ij}),$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions (see, e.g., Abramowitz and Stegun [1]). These functions, incidentally, have series representations

(14) $$\psi(r_{ij}) = \log r_{ij} - \frac{1}{2r_{ij}} - \frac{1}{12r_{ij}^2} \cdots$$

$$\psi'(r_{ij}) = \frac{1}{r_{ij}} + \frac{1}{2r_{ij}^2} + \frac{1}{6r_{ij}^3} + \ldots.$$

The maximum likelihood estimates of $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{\gamma}_{ij}$, given by expressions (8) and (9), are linear combinations of the $\log \hat{\theta}_{ij}$'s. Hence the variances and covariances of these estimates are in principle readily determined. The calculations are somewhat tedious, though if we note that the calculations required are just those required in the unbalanced two-way analysis of variance model

(15) $$Y_{ij} = \log \hat{\theta}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij},$$

where $e_{ij} \sim N(0, 1/r_{ij})$, then we can make use of standard techniques for this model (see, e.g., [6, pp. 79-91] or [8, Chapter 35]).

The statistic for testing for absence of interaction (i.e., that $\gamma_{ij} = 0$, $i = 1, \ldots, a$; $j = 1, \ldots, b$) is, in terms of expression (11), $S_0 = \underline{\hat{n}}'\, C\underline{\hat{n}}$, where $\underline{n} = (\gamma_{11}, \ldots, \gamma_{1,a-1}, \gamma_{21}, \ldots, \gamma_{(a-1)(b-1)})$. This turns out to be equal to

(16) $$S_0 = \sum_{i=1}^{a}\sum_{j=1}^{b} r_{ij}\,\hat{\gamma}_{ij}^2,$$

where $\hat{\gamma}_{ij}$ is as given in expressions (8) and (9). Asymptotically, S_0 is distributed as $\chi^2_{(a-1)(b-1)}$ if H_3 is true, and large values of S_0 give evidence against H_3.

Variances of contrasts involving the m.l.e.'s can be determined from V. In addition, tests concerning the α_i's and β_j's can also be carried out. However, tests concerning the α_i's and β_j's are unlikely to be of much interest except when the interaction terms γ_{ij} are 0. Therefore, in discussing these quantities we will suppose that we are in situations where it has been found suitable to work with the "main effects" model H_3.

The equations for obtaining the m.l.e.'s under model H_3 have already been given in expressions (10). Test and estimation procedures can be based on the limiting normal distribution of $(\hat{\mu},\hat{\alpha}_1, \ldots,\hat{\alpha}_{a-1},\hat{\beta}_1, \ldots,\hat{\beta}_{b-1})$. The covariance matrix for this limiting distribution is, as in the case of the full model, obtained from the expectations of the second derivations of log L. In this case, the covariance matrix is of the form

$$V_1 = \begin{pmatrix} r.. & \underline{0} & \underline{0} \\ \underline{0} & A & D \\ \underline{0} & D' & B \end{pmatrix}^{-1}, \tag{17}$$

where V_1 is dimension $(a + b - 1) \times (a + b - 1)$, and A, B, and D are as given in expressions (12). The matrix V_1 is readily inverted numerically in any specific application. In the proportional frequency design mentioned earlier, in which case $r_{ij} = r_{i.}\, r_{.j}/r..$, we have $D = 0$ so that

$$V_1 = \begin{pmatrix} r..^{-1} & \underline{0} & \underline{0} \\ \underline{0} & A^{-1} & 0 \\ \underline{0} & 0' & B^{-1} \end{pmatrix}. \tag{18}$$

Further, it can be shown (e.g., see Kendall and Stuart [8, p. 18]) that A^{-1} and B^{-1} have entries

$$A^{ii} = \frac{1}{r_{i.}} - \frac{1}{r..},\quad A_{ij} = -\frac{1}{r..}\ (i \neq j),$$

$$B^{ii} = \frac{1}{r_{.i}} - \frac{1}{r..},\quad B_{ij} = -\frac{1}{r..}\ (i \neq j).$$

In the general case (nonproportional frequencies), V_1 will have the form

$$V_1 = \begin{pmatrix} r..^{-1} & \underline{0} & \underline{0} \\ \underline{0} & E & F \\ \underline{0} & F' & G \end{pmatrix}. \tag{19}$$

Tests of the models H_1 or H_2 vs H_3 can be carried out by employing the limiting normal distribution of the m.l.e.'s. To test H_2: $\log \theta_{ij} = \mu + \beta_j$ vs H_3 (that is, we are testing $\alpha_i = 0$ within the model H_3), we consider the statistic

$$S_2 = \underline{\hat{\alpha}}'\, E^{-1}\, \underline{\hat{\alpha}},$$

where $\underline{\hat{\alpha}} = (\hat{\alpha}_1, \ldots,\hat{\alpha}_{a-1})$ is the m.l.e. of $\underline{\alpha}$ under H_3. If H_2 is true, S_2 is approximately $\chi^2_{(a-1)}$. Similarly, to test H_1 vs H_3 $(\beta_j = 0)$, we consider

$$S_1 = \underline{\hat{\beta}}'\, G^{-1}\, \underline{\hat{\beta}}$$

which, if H_1 is true, is approximately $\chi^2_{(b-1)}$.

Contrasts involving the α_i's or β_j's are also easily handled. If we are interested, say, in the contrast $\phi = l_1\alpha_1 + \ldots\ l_{a-1}\alpha_{a-1}$, then we consider the approximate normal distribution of $\hat{\phi} = l_1\hat{\alpha}_1 + \ldots\ l_{a-1}\hat{\alpha}_{a-1}$, which is easily obtained from the approximate joint normal distribution of $(\hat{\alpha}_1, \ldots,\hat{\alpha}_{a-1})$. In particular, $\hat{\phi}$ has a limiting normal distribution with mean ϕ and variance $\underline{l}'\, E\, \underline{l}$, where $\underline{l} = (l_1, \ldots,l_{a-1})'$.

A final remark is that, since $\mathrm{Var}(\log \hat{\theta}_{ij})$ is exactly equal to $\psi'(r_{ij})$ and only asymptotically equal to $1/r_{ij}$, it may be preferable to use $\psi'(r_{ij})$ instead of r_{ij} in the above calculations, unless the r_{ij}'s are fairly large. In this case V will be the exact covariance matrix for the estimates, and not just an approximation, although this does not of course imply that S_0 will necessarily be

more closely approximated by a χ^2 distribution than before. In the case of V_1 use of $\psi'(r_{ij})$ does not give the exact covariance matrix of the m.l.e.'s, so it is doubtful whether the above modification is worth considering.

5. LEAST SQUARES ANALYSIS

Least squares methods can be conveniently used with the types of models discussed here. We consider this only briefly, since least squares analysis is essentially the same as the approximate normal analysis using the model (15) described in Section 4.

We consider the distribution of $\log \hat{\theta}_{ij}$, and write

$$\log \hat{\theta}_{ij} = \log \theta_{ij} - \log r_{ij} + \epsilon'_{ij} \tag{20}$$

where ϵ'_{ij} has a log gamma distribution with mean $\psi(r_{ij})$ and variance $\psi'(r_{ij})$. The model (20) can be rewritten as

$$Y_{ij} = \log \theta_{ij} + \epsilon_{ij}, \tag{21}$$

where $Y_{ij} = \log \hat{\theta}_{ij} + \log r_{ij} - \psi(r_{ij})$ and $\epsilon_{ij} = \epsilon'_{ij} - \psi(r_{ij})$. In view of expressions (14) we note that

$$Y_{ij} = \log \hat{\theta}_{ij} + \frac{1}{2r_{ij}} + \frac{1}{12r_{ij}^2} + \dots$$

and ϵ_{ij} has mean 0 and variance $\psi'(r_{ij}) = 1/r_{ij} + 1/2r_{ij}^2 + \dots$. Least squares analysis is carried out by considering the model (21) and using weighted least squares with weights $w_{ij} = 1/\psi'(r_{ij})$. Tests of hypotheses can be carried out by treating the ϵ_{ij}'s as being approximately normally distributed. We observed that this analysis is similar to the approximate normal theory analysis for model H_4 described in Section 4, except that r_{ij} is replaced by w_{ij}, and $\log \hat{\theta}_{ij}$ is replaced by $\log \hat{\theta}_{ij} + \log r_{ij} - \psi(r_{ij})$. The covariance matrix for the least squares estimates of $\mu, \alpha_1, \dots, \alpha_{a-1}, \beta_1, \dots, \beta_{b-1}, \gamma_{11}, \dots, \gamma_{(a-1)(b-1)}$ is given by expression (11), except that r_{ij} is replaced by w_{ij} in Eqs. (12). This method and the approximate normal methods of Section 4 do not differ by much, unless the r_{ij}'s are fairly small, since to first order (14), $\psi'(r_{ij}) = 1/r_{ij}$.

For more details on weighted least squares in the two-way model, we refer the reader as before to Kempthorne [6, pp. 79-91] or Kendall and Stuart [8, Ch. 35].

6. AN EXAMPLE

In handling data from the model (5), we have two main choices with regard to overall test of models H_s vs H_t: one choice is to use likelihood ratio tests, the other to use tests based on the approximate normality of the m.l.e.'s or least squares estimates of the parameters. Both methods require a moderate amount of computation. For example, in using likelihood ratio statistics we may need to solve equations iteratively to obtain m.l.e.'s. If we use least squares estimates, this is not necessary, though we do have to solve linear equations for the estimates. However, when we use a statistic such as S_1 or S_2, we need to invert matrices, which is not required for the likelihood ratio tests. In order to estimate contrasts for parameters, maximum likelihood or least squares methods can be used. In both cases it is necessary to obtain the covariance matrices V or V_1, and to invert matrices. The least squares estimates are slightly more easily calculated, if model H_3 is used, though they are also slightly less efficient than the m.l.e.'s (see Section 7).

We will now demonstrate the use of the procedures in an example, and return briefly to a further discussion of the merits of the procedures in Section 7.

EXAMPLE: We consider for convenience a simulated 3×3 factorial experiment with $\theta_{ij} = \exp(2i + 2j)$, $i = 0, 1, 2$; $j = 0, 1, 2$. The values of T_{ij} and r_{ij} (in brackets) are given below:

	i \ j	B 0		1		2		$T_{i.}\ (r_{i.})$	
	0	9.61	(10)	48.14	(6)	102.7	(4)	160.45	(20)
A	1	46.88	(6)	281.7	(4)	463.7	(2)	792.28	(12)
	2	254.7	(4)	807.8	(2)	8431.0	(2)	9493.50	(8)
$T_{.j}$	$(r_{.j})$	311.19	(20)	1137.64	(12)	8997.4	(8)	10446.23	(40)

We begin by testing for absence of the interaction term in the model $\log \theta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$. Using the likelihood ratio test of H_3 vs H_4, we find $\lambda_{34} = 2(\Lambda_3 - \Lambda_4) = 1.227$. Comparing this with percentage points of the $\chi^2_{(4)}$ distribution, we see there is no evidence against H_3, and hence in further discussions we will work with model H_3. We note that in order to calculate λ_{34} we had to compute the m.l.e. $\hat{\mu}$ under model H_4 (see Eqs. (8)), and also the m.l.e.'s of μ, α_i, β_j under H_3. These latter quantities are found by solving Eqs. (10) to be $\hat{\mu} = 2.709$, $\hat{\alpha}_1 = -1.521$, $\hat{\alpha}_2 = 0.625$, $\hat{\alpha}_3 = 2.864$, $\hat{\beta}_1 = -1.352$, $\hat{\beta}_2 = 0.767$, and $\hat{\beta}_3 = 2.228$.

Alternately, we could test H_3 vs H_4 by using the large sample approximate normal distributions of either the m.l.e.'s or the least squares estimates. Using S_0 given by Eq. (16), we find $S_0 = 1.207$, which is in close agreement with the likelihood ratio test result. Using S_0, but using the least squares estimates and their variances instead of the m.l.e.'s, gives $S_0 = 0.935$, in good agreement with the other results. The least squares estimates of the parameters under model H_3 are, incidentally, $\tilde{\mu} = 2.747$, $\tilde{\alpha}_1 = -1.476$, $\tilde{\alpha}_2 = 0.726$, $\tilde{\alpha}_3 = 2.972$, $\tilde{\beta}_1 = -1.316$, $\tilde{\beta}_2 = 0.851$, and $\tilde{\beta}_3 = 2.323$.

Effects within the model H_3 can be examined further using either the m.l.e.'s or least squares estimates. We will illustrate here the use of the m.l.e.'s; as noted above, least squares methods are computationally similar to these. The m.l.e.'s of the parameters have been given above. We also require the covariance matrix V_1^{-1}, which is readily found from expressions (17) and (12) to be

$$V_1 = \begin{vmatrix} 0.0250 & 0 & 0 & 0 & 0 \\ 0 & 0.0250 & -0.0250 & 0.0 & 0.0 \\ 0 & -0.0250 & 0.0586 & 0.0 & -0.0028 \\ 0 & 0.0 & 0.0 & 0.0250 & -0.0250 \\ 0 & 0.0 & -0.0028 & -0.0250 & 0.0586 \end{vmatrix}.$$

Overall tests of the main effects A and B can be carried out using either statistics S_1 and S_2, or the likelihood ratio tests. The former requires the inversion of some 2×2 matrices, but does not require any new estimates to be calculated. The latter method requires us to calculate new m.l.e.'s under H_2 or H_1, but no matrix inversion. In this and similar problems, the amount of computation required by the two methods is comparable, and the two methods will usually give results in fairly close agreement, so it is simply a matter of preference which procedure one uses.

Using the likelihood ratio test for H_2 vs H_3, we find $\lambda_{23} = 108.9$, and for H_1 vs H_3 we find $\lambda_{13} = 74.6$. The corresponding values of S_1 and S_2 are 116.4 and 83.3; use of $\psi'(r_{ij})$

instead of $1/r_{ij}$ in V_1, as discussed at the end of Section 4, changes these only marginally, to 117.2 and 83.4. The upper 0.01 percent point $\chi^2_{(2)}$ is 9.21, so that both tests show strong evidence in favor of main effects for both A and B.

If desired, contrasts involving the parameters can also be examined. For example, in the present situation it would be of interest to consider constrasts which correspond to linear and quadratic effects in A and B. For example, for A we might consider $\phi_{LA} = \alpha_3 - \alpha_1$ and $\phi_{QA} = \alpha_3 - 2\alpha_2 + \alpha_1$. Estimates of these contrasts and their variances are easily obtained from $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\alpha}_3$, and V_1, and are as follows:

$$\hat{\phi}_{LA} = \hat{\alpha}_3 - \hat{\alpha}_1 = 4.385, \quad \hat{Var}(\hat{\phi}_{LA}) = 0.1755,$$

$$\hat{\phi}_{QA} = \hat{\alpha}_3 - 2\hat{\alpha}_2 + \hat{\alpha}_1 = 0.0931, \quad \hat{Var}(\hat{\phi}_{QA}) = 0.1755.$$

In calculating variances of $\hat{\phi}_{LA}$ and $\hat{\phi}_{QA}$, we make use of the restriction $20\alpha_1 + 12\alpha_2 + 8\alpha_3 = 0$ to rewrite ϕ_{LA} as $-3.5\hat{\alpha}_1 - 1.5\hat{\alpha}_2$ and $\hat{\phi}_{QA}$ as $-1.5\hat{\alpha}_1 - 3.5\hat{\alpha}_2$. Tests of $\phi_{LA} = 0$ and $\phi_{LA} = 0$ can be carried out by treating $\hat{\phi}^2_{LA}/\hat{Var}(\phi_{LA})$ and $\hat{\phi}^2_{QA}/\hat{Var}(\phi_{QA})$ as approximately $\chi^2_{(1)}$. Here we find, that $\hat{\phi}^2_{LA}/\hat{Var}(\phi_{LA}) = 109.5$, and $\hat{\phi}^2_{QA}/\hat{Var}(\hat{\phi}_{QA}) = 0.017$, indicating very strong evidence in favor of a linear, but no quadratic, effect on $\log \theta$ for factor A. (This is of course as we expect, since the simulated data come from the model $\log \theta_{ij} = 2i + 2j$.)

In a model in which there is no interaction effect, it may also be of interest to estimate, say, $\alpha_i - \alpha_j$. Note that $\alpha_i - \alpha_j = \log \theta_{ik} - \log \theta_{jk} = \log (\theta_{ik}/\theta_{jk})$ represents the log of the ratio of the mean lifetime of items at levels i and j of factor A (and the same level of factor B). For example, above we have found that $\hat{\alpha}_3 - \hat{\alpha}_1 = 4.385$; if we desire an approximate confidence interval for $\alpha_3 - \alpha_1$, we find it using $\hat{Var}(\hat{\phi}_{LA}) = 0.1755$, so that an approximate 0.95 confidence interval for $\alpha_3 - \alpha_1$ is given by $\hat{\phi}_{LA} \pm 1.96\sqrt{\hat{Var}(\hat{\phi}_{LA})}$. In this case, this yields the interval $3.564 \leqslant \alpha_3 - \alpha_1 \leqslant 5.206$. This gives confidence limits on the ratio θ_{3k}/θ_{1k} of mean lifetimes of items at levels 1 and 3 of A as $35.30 \leqslant \theta_{3k}/\theta_{1k} \leqslant 182.4$.

7. ADDITIONAL REMARKS

The procedures described are all easy to use. The least squares procedures do not require any iterative solutions for estimates, though with an unbalanced design the computation required for the least squares analysis is not much less than that required by the maximum likelihood or likelihood ratio procedures. The maximum likelihood methods are, in experiments having many factor combinations, more efficient than the least squares procedures, especially when the r_{ij}'s are not large [3]. All the procedures described rely, however, on large sample approximations, and so the question arises as to how well various statistics are approximated by their limiting χ^2 or normal distributions. The log gamma distribution of $\hat{\theta}_{ij}$ approaches normality fairly rapidly as r_{ij} increases, so that with even moderate sized r_{ij}'s (say about 8 or more), the approximations employed should be sufficiently accurate for practical purposes. Zelen [16] has performed a few simulations which support this, though unfortunately some of his results refer to an incorrect likelihood ratio statistic. For even smaller values of r_{ij}, the approximations should be reasonably accurate, though in this case χ^2 approximations involved with likelihood ratio tests might be expected to be slightly better than those involved with the distributions of m.l.e.'s or least squares estimates.

We also remark that although we have restricted our discussion to a two-way model, more general analysis-of-variance or regression models can be handled in a similar way. Hence the large number of models which are of the form (1), with $\log \theta$ a linear function of fixed effects or covariables, can, for example, be readily handled.

Finally, no work on the exponential distribution is complete without a reminder that procedures based on the exponential model are rather nonrobust. That is, they are sensitive to departures from exponentiality. This is equally true of the maximum likelihood and the least squares procedures. Therefore, in using procedures such as those discussed in this paper, checks need to be made on the adequacy of the assumed exponential model.

REFERENCES

[1] Abramowitz, M. and I.A. Stegun, *Handbook of Mathematical Functions*, Dover, New York (1965).

[2] Bartlett, M.S. and D.G. Kendall, "The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation," Journal of the Royal Statistical Society, Suppl. *8*: 128-138 (1946).

[3] Cox, D.R. and D.V. Hinkley, "A Note on the Efficiency of Least Squares Estimates," Journal of the Royal Statistical Society B, *30*: 284-289 (1968).

[4] Epstein, B. and M. Sobel, "Life Testing," Journal of the American Statistical Association *48*: 486-502 (1953).

[5] Kahn, H.D., "Least Squares Analysis of Accelerated Life Tests for the Power Rule and Arrhenius Models," in *The Theory and Applications of Reliability*, C.P. Tsokos and I.N. Shimi (Eds.) Vol. II, (7): 437-456 Academic Press, New York, (1977).

[6] Kempthorne, O., *The Design and Analysis of Experiments*, Wiley, New York (1952).

[7] Kendall, M.G. and A. Stuart, *The Advanced Theory of Statistics, Vol. II*, Charles Griffin and Co., London (1967).

[8] Kendall, M.G., *The Advanced Theory of Statistics, Vol. III*, Charles Griffin and Co., London (1968).

[9] Lawless, J.F., "Confidence Interval Estimation in the Inverse Power Law Model," Applied Statistics *25*: 128-138 (1976).

[10] Lawless, J.F. and K. Singhal, "Screening of Nonnormal Regression Models," Biometrics 34:318-27 (1978).

[11] Mann, N.R., R. Schafer and N.D. Singpurwalla, *Methods for Statistical Analysis of Reliability and Life Data*, Wiley, New York (1974).

[12] Nelson, W.B., "Statistical Methods for Accelerated Life Test Data-the Inverse Power Law Model," General Electric Co. Tech. Report No. 71-C-011 (1970).

[13] Prentice, R.L., "Exponential Survivals with Censoring and Explanatory Variables," Biometrika *60*: 279-288 (1973).

[14] Singpurwalla, N.D., "A Problem in Accelerated Like Testing," Journal of the American Statistical Association *66*: 841-845 (1971).

[15] Singpurwalla, N.D., V.F. Castellino and D.Y. Goldschen, "Inference from Accelerated Life Testing Using Eyring Type Reparameterizations," Naval Research Logistics Quarterly *22*: 289-296 (1973).

[16] Zelen, M., "Factorial Experiments in Life Testing," Technometrics *1*: 269-288 (1959).

[17] Zelen, M., "Analysis of Two-Factor Classifications with Respect to Life Tests," in *Contributions to Statistics*, I. Olkin (Ed.) Stanford University Press, San Francisco (1960).

ESTIMATING THE ECONOMIC IMPACT OF THE 1973 NAVY BASE CLOSING: MODELS, TESTS, AND AN *EX POST* EVALUATION OF THE FORECASTING PERFORMANCE

Randolph F. C. Shen

Department of Management Science
University of Rhode Island
Kingston, Rhode Island

ABSTRACT

In 1973 the Defense Department made plans to close many Navy bases in the United States. Hardest hit was Rhode Island which would suffer a loss of 45.61% of the total cutback of 42,812 jobs. This paper describes two models which were built to forecast the severity of the economic impact in Rhode Island: one used the reduced form equation approach, and the other the simultaneous equations system approach. Tests on multicollinearity, specification, and serial correlation were conducted. An *ex post* evaluation of these two models' performance in forecasting then concludes the paper.

1. INTRODUCTION

On April 16, 1973, the Defense Department announced plans to close or modify installations in 30 states, the District of Columbia, and Puerto Rico. It would eliminate 26,172 civilian and 16,640 military jobs by June 30 of the next year. Hardest hit by this cutback was Rhode Island, whose Quonset Point Naval Air Station would be shut down with a loss of 3,809 civilian and 4,217 military jobs, and whose Newport Naval Base would be severely cut down, eliminating 430 civilian and 11,069 military positions. This meant that Rhode Island would suffer a primary job loss of 4,239 civilian and 15,286 military employments with a total of 19,525, which was 45.61% of the total cutback of 42,812 jobs. This announcement was greeted in Rhode Island with "shock, anger, and bewilderment," and was ranked to be as catastrophic as "the disastrous hurricanes" experienced by the state in the past.

A study was conducted in the summer of that year to analyze, to assess, and to forecast the severity of the economic impact of the Navy base closing in Rhode Island. The results were reported as a headline story in *The Providence Sunday Journal* [2] on August 5, 1973.

Two forecasting models were built for this economic impact study: one for the employment and the other for the retail sales subject to taxation. However, only the employment model will be discussed here.

2. IMPACT MULTIPLIER, REDUCED FORM EQUATION, AND SINGLE EQUATION LEAST SQUARES METHOD

2.1 Specification of the Model

In the employment model, a five-equation system is built as follows: the first equation is an identity (Eq. (2.1)), which defines the total nonagricultural wage-and-salaried employment (T) in the state as consisting of manufacturing (M) and nonmanufacturing (N) employment; the latter includes Navy civilian workers (C) as a part of its government employment, but not Navy military personnel (D). Thus, nonmanufacturing (N) employment is equal to Navy civilian workers (C) plus the remaining, which we shall call service (S) employment:

$$T = M + N = M + S + C, \tag{2.1}$$

where $N = S + C$.

Based upon past research, it has been found that Rhode Island manufacturing (M) employment is strongly subject to cyclical influence but her nonmanufacturing (N) employment is not. Therefore, it is hypothesized that in Eq. (2.2) manufacturing (M) employment is a function of both time trend (t) and a cyclical dummy variable (V), using 1 as downswing and 0 as upswing:

$$M = a + b\,t + c\,V + u_2. \tag{2.2}$$

In Eq. (2.3), service (S) employment is assumed, following the economic base theory, to serve on a supportive basis for the state's total (T) employment and Navy military personnel (D). Thus, we have

$$S = h + k\,T + w\,D + u_3. \tag{2.3}$$

Finally, Navy civilian workers (C) and military personnel (D) are taken to be exogenous as they are specified in Eqs. (2.4) and (2.5):

$$C = \bar{C}, \tag{2.4}$$

$$D = \bar{D}. \tag{2.5}$$

2.2 Estimation, Testing, and Evaluation

Since the immediate interest back in 1973 was the impact effect of the Navy base closing, Eqs. (2.1) and (2.2) were substituted into Eq. (2.3) and solved for S in terms of t, V, C and D:

$$S = \frac{h + ka}{1 - k} + \frac{kb}{1 - k}\,t + \frac{kc}{1 - k}\,V + \frac{k}{1 - k}\,C + \frac{w}{1 - k}\,D + \frac{ku_2 + u_3}{1 - k}. \tag{2.6}$$

Using time series data from 1961 to 1972, this reduced form equation for the service (S) employment was estimated by the single equation least squares method. It turns out to have an annual growth factor of 5.647, and a recession coefficient of 3.928. The Navy civilian and military employment impact multipliers are estimated to be 0.499 and 0.216, respectively. All these statistics are presented in Table 1.

The "t" test for time trend (t) is found to be statistically significant at more than 1% level; for cyclical effect (V), significant at 5% level; for Navy military personnel (D), significant at about 30% level; and for Navy civilian workers (C), not significant at all. The $\bar{R}^2$ is 0.993, and the calculated F with 4 and 7 degrees of freedom is 392.721.

TABLE 1 — *Estimation of Impact Multiplier and Reduced Form Equation*

Service Employment Equation*		
Variable	Estimated Coefficient	*t* statistic
Y-intercept	147.444	—
Time	5.647	12.915
One, if a downswing	−3.928	−2.645
Navy civilian workers	0.499	0.311
Navy military personnel	0.216	1.018

*$\bar{R}^2 = 0.993$, $F(4, 7) = 392.721$, $D.W. = 2.049$, $S.E.E. = 1.639$.

To compare the *ex ante* forecast which was reported by Goodrich [2] against the *ex post* forecast, Tables 2 and 3 are prepared.

TABLE 2 — *Ex Ante Forecasting of Secondary Job Loss Based on Planned Reduction*

Year	Category	Planned Reduction Civilian Workers	Planned Reduction Military Personnel	Multiplier	Secondary Job Loss
1973	Civilian Workers	2,518*		0.499	−1,257
	Military Personnel		12,244†	0.216	−2,645
Total		2,518	12,244	—	−3,902
1974	Civilian Workers	1,721*		0.499	−859
	Military Personnel		3,042†	0.216	−657
Total		1,721	3,042	—	−1,516

*Total *planned* reduction in Navy civilian workers in 1973 and 1974 is distributed according to the percentage calculated from the *actual* reduction in 1973 and 1974.
†Same as above except it is related to Navy military personnel.

TABLE 3 — *Ex Post Forecasting of Secondary Job Loss Based on Actual Reduction*

Year	Category	Actual Reduction Civilian Workers	Actual Reduction Military Personnel	Multiplier	Secondary Job Loss
1973	Civilian Workers	3,504		0.499	−1,749
	Military Personnel		18,523	0.216	−4,001
Total		3,504	18,523	—	−5,750
1974	Civilian Workers	2,395		0.499	−1,195
	Military Personnel		4,609	0.216	−996
Total		2,395	4,609	—	−2,191

It is interesting to note that first, the *actual* reduction in civilian workers and military personnel by the Navy is greater than the announced *planned* reduction; and, consequently, the *ex ante* forecast of the secondary job loss is smaller than the *ex post* forecast. This shows once again that the conditional forecasting is different from the unconditional forecasting. Second, among the many other forecasts made in 1973, our *ex ante* forecast turned out to be the most optimistic one in the sense that the economic impact of the defense base closing would be the least severe. Had our forecast used the actual reduction figure, it would have sounded more alarm. But events proved to be the contrary. For instance, in 1973 the secondary job loss should have been forecast to be 5,750 based on the actual reduction figure (not 3,902 based on the planned reduction figure), and the *ex post* forecast for Rhode Island total employment should have been 356,351; but the actual observed total employment was 365,900, and the underestimation was 9,549. Thus, we have to conclude that either Rhode Island's inner economic strength is actually stronger, or the economic impact of the defense base closing is much weaker than suspected, or both.

Since a high degree of multicollinearity implies that at least one of the explanatory variables is a linear function of one or more of the remaining explanatory variables plus residual, to run the test for multicollinearity in our sample, we regress each of the explanatory variables on all the remaining explanatory variables. The functional forms and the $\bar{R}^2$ values are listed in Table 4. The highest $\bar{R}^2$ is 0.865, which may be taken as a measure of the multicollinearity in our sample.

TABLE 4 — *Multicollinearity Test for the Service Employment Equation*

Functional Form	$\bar{R}^2$
$C = g(t, V, D)$	0.814
$D = h(t, V, C)$	0.344
$t = i(V, C, D)$	0.865
$V = j(t, C, D)$	0.005

TABLE 5 — *Evaluation of the Forecasting Performance of the Reduced Form Equation*

Equation	Year	Forecast Value	Actual Value	Error	Percent Error
Service Employment	1973	225.642	235.191	−9.549	−0.041
	1974	225.171	237.686	−12.515	−0.053
	1975	230.918	232.270	−1.352	−0.006
	1976	240.414	239.737	+0.677	+0.003
	Mean Absolute Error			6.023	—
	Mean Absolute Percent Error			—	0.026
Manufacturing Employment	Not Estimated				
Total Employment	1973	356.351	365.900	−9.549	−0.026
	1974	354.185	366.700	−12.515	−0.034
	1975	347.848	349.200	−1.352	−0.004
	1976	366.977	366.300	+0.677	+0.002
	Mean Absolute Error			6.023	—
	Mean Absolute Percent Error			—	0.016

The specification error tests are conducted by distinguishing among various regression models. Many regressions with different specifications are run. Those retained here seem to have the correct sign and magnitude.

The Durbin-Watson statistic is calculated to be 2.049, which indicates no serial correlation in this service employment equation.

Lastly, in Table 5 we evaluate the forecasting performance of our estimation by calculating the difference between the actually observed value and the forecast value using the *ex post* data for the explanatory variables. This yields a mean absolute error of 6.023 and a mean absolute percent error of 2.6% for the service employment equation, whereas for the total employment equation the mean absolute error is 6.023, and the mean absolute percent error is 1.6%.

3. IMPACT AND LAGGED MULTIPLIERS, SIMULTANEOUS EQUATIONS SYSTEM, AND TWO-STAGE LEAST SQUARES METHOD

3.1 Specification of the Model

More recently, a simultaneous equations system has been tested with a partial adjustment model built in as its theoretical basis. Assuming the *desired* level of the manufacturing employment (M^*) in the state in a year is a linear function of time trend (t), cyclical influence (V), and cyclical influence lagged by one year (V_{-1}), we have

$$M^* = a + b\ t + c\ V + d\ V_{-1} + u_1. \tag{3.1}$$

Due to institutional rigidity, technological constraints, and other reasons, it is theorized that the actual level and the desired level can only be adjusted partially as follows:

$$M - M_{-1} = k(M^* - M_{-1}) + \mathrm{v}, \tag{3.2}$$

where $0 \leqslant k \leqslant 1$, and v is a random disturbance.

Solving for M^* in Eq. (3.2) and inserting the result into Eq. (3.1), we obtain

$$M = ak + bk\ t + ck\ V + dk\ V_{-1} + (1-k)\ M_{-1} + (ku_1 + \mathrm{v}). \tag{3.3}$$

Similarly, we work out the service employment equation as the following:

$$S^* = h + s(M + C) + m(M + C)_{-1} + w\ D + r\ t + u_4, \tag{3.4}$$

$$S - S_{-1} = g(S^* - S_{-1}) + e, \tag{3.5}$$

$$S = hg + sg(M + C) + mg(M + C)_{-1} + wg\ D + rg\ t + (1 - g)\ S_{-1} + (gu_4 + e). \tag{3.6}$$

Equations (3.3) and (3.6) together with the identity (2.1) constitute our second model, which is a simultaneous equations system and which will be estimated by the two-stage least squares method.

3.2 Estimation, Testing, and Evaluation

Since all the explanatory variables on the right-hand side of Eq. (3.3) are either exogenous or predetermined, we, using the same time series data but one year less, estimate Eq. (3.3) first by the least squares method. The result is presented in Table 6(a).

TABLE 6 — *Estimation of Impact and Lagged Multipliers and Simultaneous Equations System*

(a) Manufacturing Employment Equation*		
Variable	Estimated Coefficient	*t* Statistic
Y-intercept	60.225	—
Time	0.952	1.516
One, if a downswing	−8.351	−3.065
One, if a downswing lagged 1 yr.	−5.487	−1.033
Manufacturing employment lagged 1 yr.	0.471	1.483

*$\bar{R}^2 = 0.707$, $F(4, 6) = 7.023$, *S.E.E.* = 2.831.

(b) Service Employment Equation*		
Variable	Estimated Coefficient	*t* Statistic
Y-intercept	125.533	—
Calculated manufacturing employment and Navy civilian workers	0.377	1.580
Manufacturing employment and Navy civilian workers lagged 1 yr.	−0.237	−0.977
Navy military personnel	0.098	0.317
Time	5.105	1.460
Service employment lagged 1 yr.	0.078	0.112

*$\bar{R}^2 = 0.987$, $F(5, 5) = 151.535$, *S.E.E.* = 2.089.

In Eq. (3.6), M is the only endogenous variable on the right-hand side of the equation, so we use the estimated values of M and estimate Eq. (3.6) by the least squares method again. This produces the statistics in Table 6(b).

Finally, using the estimated regression coefficients obtained from Eqs. (2.1), (3.3), and (3.6), we form the following matrix equation:

$$\underset{\sim}{W}\underset{\sim}{Y}_t = \underset{\sim}{A}\underset{\sim}{Y}_{t-1} + \underset{\sim}{B}\underset{\sim}{X}_t + \underset{\sim}{C}\underset{\sim}{X}_{t-1} + \underset{\sim}{u}_t, \tag{3.7}$$

where $\underset{\sim}{W}$, $\underset{\sim}{A}$, $\underset{\sim}{B}$, and $\underset{\sim}{C}$ are estimated regression coefficient matrices, $\underset{\sim}{Y}_t$ and $\underset{\sim}{Y}_{t-1}$ are the column vectors of the endogenous variables, and $\underset{\sim}{X}_t$ and $\underset{\sim}{X}_{t-1}$ the column vectors of the exogenous variables, at periods t and $t-1$, respectively.

Premultiplying Eq. (3.7) by $\underset{\sim}{W}^{-1}$, we have

$$\begin{aligned} \underset{\sim}{Y}_t &= \underset{\sim}{W}^{-1}\underset{\sim}{A}\underset{\sim}{Y}_{t-1} + \underset{\sim}{W}^{-1}\underset{\sim}{B}\underset{\sim}{X}_t + \underset{\sim}{W}^{-1}\underset{\sim}{C}\underset{\sim}{X}_{t-1} + \underset{\sim}{W}^{-1}\underset{\sim}{u}_t, \\ &= \underset{\sim}{A}^*\underset{\sim}{Y}_{t-1} + \underset{\sim}{B}^*\underset{\sim}{X}_t + \underset{\sim}{C}^*\underset{\sim}{X}_{t-1} + \underset{\sim}{W}^{-1}\underset{\sim}{u}_t, \end{aligned} \tag{3.8}$$

where $\underset{\sim}{A}^* = \underset{\sim}{W}^{-1}\underset{\sim}{A}$, $\underset{\sim}{B}^* = \underset{\sim}{W}^{-1}\underset{\sim}{B}$, and $\underset{\sim}{C}^* = \underset{\sim}{W}^{-1}\underset{\sim}{C}$.

Now, lagging the matrix Eq. (3.8) by one period and substituting the result into the same matrix equation repeatedly, we obtain

$$\begin{aligned} \underset{\sim}{Y}_t = {} & (\underset{\sim}{A}^*)^k\underset{\sim}{Y}_{t-k} + \underset{\sim}{B}^*\underset{\sim}{X}_t + (\underset{\sim}{A}^*\underset{\sim}{B}^* + \underset{\sim}{C}^*)\underset{\sim}{X}_{t-1} + \underset{\sim}{A}^*(\underset{\sim}{A}^*\underset{\sim}{B}^* + \underset{\sim}{C}^*)\underset{\sim}{X}_{t-2} \\ & + (\underset{\sim}{A}^*)^2(\underset{\sim}{A}^*\underset{\sim}{B}^* + \underset{\sim}{C}^*)\underset{\sim}{X}_{t-3} + \cdots\cdots \\ & + (\underset{\sim}{A}^*)^{k-2}(\underset{\sim}{A}^*\underset{\sim}{B}^* + \underset{\sim}{C}^*)\underset{\sim}{X}_{t-k+1} + (\underset{\sim}{A}^*)^{k-1}\underset{\sim}{C}^*\underset{\sim}{X}_{t-k}, \end{aligned} \tag{3.9}$$

where we have, of course, ignored the residual term.

In order to have the system to be stable, it is required that $\lim (\underset{\sim}{A}^*)^k = 0$ when $k \to \infty$, and all the characteristic roots of $\underset{\sim}{A}^*$ be < 1 in its absolute value. The coefficients of $\underset{\sim}{B}^*$ will then serve as our impact multiplier after one-shot increase of its related exogenous variable, the coefficients of $(\underset{\sim}{A}^* \underset{\sim}{B}^* + \underset{\sim}{C}^*)$ our lagged multiplier lagged by one period, $\underset{\sim}{A}^*(\underset{\sim}{A}^* \underset{\sim}{B}^* + \underset{\sim}{C}^*)$ our lagged multiplier lagged by two periods, etc.

In our case, the system is found to be stable, and the impact multipliers for Navy civilian worker and military personnel are estimated to be 0.377 and 0.098, respectively. When compared to the impact multipliers of 0.499 and 0.216 from the reduced form equation in the first model, it is found that they are slightly smaller, and seem more reasonable as the aftereffects are known now.

Those lagged multipliers for Navy civilian workers lagged by one, two, and three years are estimated to be 0.029, 0.0023, and 0.00018, while those for Navy military personnel are estimated at 0.008, 0.0006, and 0.000047, respectively.

The secondary job loss for 1973 and 1974 is calculated in Table 7, which should be compared to Table 3. The impact effect in 1973 turns out to be smaller, and the impact and lagged effects are also smaller in 1974. The advantage of this second model is then the capability of calculating both the impact effect as well as the lagged effect lagged as many years as required.

TABLE 7 — *Estimation of Secondary Job Loss by Impact Multiplier for 1973 and by Impact and Lagged Multipliers for 1974*

Year	Category	Actual Reduction		Multiplier	Secondary Job Loss
		Civilian Workers	Military Personnel		
1973	Civilian Workers	03,504		0.377	−1,321
	Military Personnel		18,523	0.098	−1,815
Total		3,504	18,523	—	−3,136
1974	Civilian Workers	02,395(74)		0.377	−903
		03,504(73)		0.029	−102
	Military Personnel		04,609(74)	0.098	−452
			18,523(73)	0.008	−148
Total		05,899	23,132	—	−1,605

To demonstrate the lagged effects for many years, we calculate the cumulated secondary job loss for the year of 1976 in Table 8.

Following the same test for multicollinearity as we did before, we regress each of the explanatory variables on all the remaining explanatory variables for both the manufacturing employment equation and the service employment equation. The results are collected in Table 9, (a) and (b). It seems that the degree of multicollinearity is high in the manufacturing employment equation (the highest $\bar{R}^2 = 0.867$), and that in the service employment equation it is even higher (the highest $\bar{R}^2 = 0.998$). This is so because of the fact that time trend (t) plays a more significant role in the latter equation than in the former equation.

To test the specification error, we run many regressions with different specifications again. The estimated Eqs. (3.3) and (3.6) again seem to have the correct sign and magnitude.

Next, we run the special test for serial correlation as given by Durbin [1], as in our least squares regression some of the regressors are lagged dependent variables. However, the h

TABLE 8 — *Estimation of Secondary Job Loss by Impact and Lagged Multipliers for 1976*

Category	Actual Reduction		Multiplier	Secondary Job Loss
	Civilian Workers	Military Personnel		
	67(76)		0.3771	−25
Civilian	+216(75)		0.2943E-1	+6
Workers	2,395(74)		0.2296E-2	−5
	3,504(73)		0.1791E-3	−1
		211(76)	0.9822E-1	−21
Military		34(75)	0.7663E-2	−0
Personnel		4,609(74)	0.5979E-3	−3
		18,523(73)	0.4665E-4	−1
Total	5,750	23,377	—	−50

TABLE 9 — *Multicollinearity Tests for the (a) Manufacturing and (b) Service Employment Equations*

Functional Form For (a)	$\bar{R}^2$
$t = g(V, V_{-1}, M_{-1})$	0.858
$V = h(t, V_{-1}, M_{-1})$	0.240
$V_{-1} = i(t, V, M_{-1})$	0.867
$M_{-1} = j(t, V, V_{-1})$	0.781
Functional Form For (b)	$\bar{R}^2$
$M_c + C = g((M + C)_{-1}, D, t, S_{-1})$	0.720
$(M + C)_{-1} = h((M_c + C), D, t, S_{-1})$	0.801
$t = i((M_c + C), (M + C)_{-1}, D, S_{-1})$	0.997
$D = j((M_c + C), (M + C)_{-1}, t, S_{-1})$	0.452
$S_{-1} = k((M_c + C), (M + C)_{-1}, D, t)$	0.998

statistic cannot be calculated because $n\hat{V}(b_1) > 1$, where b_1 is the regression coefficient of the lagged dependent variable. Following the procedure given at the same source, we run the regression of e_t (the residual at period t) on e_{t-1}, t, V, V_{-1}, and M_{-1} for the manufacturing employment equation, and find that the t statistic for e_{-1} is not significant at the 1% level. However, when we run the regression of e_t on e_{t-1}, $(M_c + C)$, $(M + C)_{-1}$, D, t, and S_{-1} for the service employment equation, the t statistic for e_{t-1} is significant at the 5% level but not at the 1% level. Thus, it is concluded that serial correlation is not present in the manufacturing employment equation, but is suspected in the service employment equation.

Lastly, the forecasting performance of our second model is evaluated in Table 10. The mean absolute errors for the service, manufacturing, and total employment equations are found to be 5.059, 3.333, and 6.133, and the mean absolute percent errors are 0.021, 0.028, and 0.017, respectively. They are extremely small again.

TABLE 10 — *Evaluation of the Forecasting Performance of the Simultaneous Equations System*

Equation	Year	Forecast Value	Actual Value	Error	Percent Error
Service Employment	1973	230.672	235.191	−4.519	−0.019
	1974	232.798	237.686	−4.888	−0.021
	1975	236.991	232.270	+4.721	+0.020
	1976	245.846	239.737	+6.109	+0.025
	Mean Absolute Error			5.059	—
	Mean Absolute Percent Error			—	0.021
Manufacturing Employment	1973	128.886	124.300	+4.586	+0.037
	1974	123.747	125.000	−1.253	−0.010
	1975	119.542	112.700	+6.842	+0.061
	1976	123.052	122.400	+0.652	+0.005
	Mean Absolute Error			3.333	—
	Mean Absolute Percent Error			—	0.028
Total Employment	1973	365.967	365.900	+0.067	+0.000
	1974	360.559	366.700	−6.141	−0.017
	1975	360.763	349.200	+11.563	+0.033
	1976	373.061	366.300	+6.761	+0.018
	Mean Absolute Error			6.133	—
	Mean Absolute Percent Error			—	0.017

4. CONCLUSIONS

Based upon the preceding study, it is concluded that

- An *ad hoc* economic impact study should be conducted for the state or region involved whenever there is a defense base closing. The data, variables and models should be directly related to the peculiar economic and demographic conditions of the state or region. The current practice by the Defense Department of using an employment multiplier of 2.585 for the civilian workers and an employment multiplier of 0.662 for the military personnel, which were estimated from a cross-section of data of 15 defense base closings, should be avoided (see Lynch [4]). In our 1973 case, there is no doubt now about the very unrealistic forecasts made from these employment multipliers.

- If resources are in short supply, a simple reduced form equation which gives the impact multiplier only may be worked out; otherwise, a simultaneous equations system should be estimated. The latter presents not only the impact multiplier but also the lagged multipliers lagged as many years as desired.

- It is often criticized that the above study contains a large measure of multicollinearity which renders the research useless. As it is argued by Kmenta [3], multicollinearity is "a question of degree and not of kind." In most of the real world data, it seems that there is always some degree of multicollinearity present in the sample. As long as our experiments cannot be controlled, multicollinearity seems to be unavoidable in any social or economic environment. Further, since our purpose is focused on forecasting rather than explanation, it seems that as long as the forecast is accurate, multicollinearity is of secondary importance.

• Further research may be carried out by disaggregating the manufacturing and service employment into various two-digit SIC groups or some particular industries in the state, which are closely related to the defense base closings, if desired.

• If time, financial resources, and data at the state level are available, a more sophisticated model of various employments may, based upon the neoclassical theory of the firm, be specified by introducing, say, the constant-elasticity-of-substitution (CES) production function with efficiency, distribution, substitution, and return-of-scale parameters included, if appropriate. Then, another question arises: in view of the high forecasting performance of this simple model, is the benefit gained by economic sophistication worth the cost; or, in other words, will the economically sophisticated model perform better?

ACKNOWLEDGMENT

The author wishes to express his gratitude to Mr. Joseph L. Goodrich, financial editor of *The Providence Journal-Bulletin*, for his immense assistance in supplying the data from the U.S. Navy [5] and for his discerning insight into the true nature of the problem.

REFERENCES

[1] Durbin, J., "Testing for Serial Correlation in Least Squares Regression When Some of the Regressors are Lagged Dependent Variables," *Econometrica 38*, 410-21 (1970).
[2] Goodrich, J.L., "R.I. Seen Able to Absorb Base Closings," *The Providence Sunday Journal 89*, p. 1, p. B4, August 5, 1973.
[3] Kmenta, J., *Elements of Econometrics*, The Macmillan Co., N.Y., (1971).
[4] Lynch, J.E., *Local Economic Development After Military Base Closures*, Praeger, N.Y., pp. 261-78 (1970).
[5] U.S. Navy, *Annual Report by the Navy in the Rhode Island Area*, Newport, R.I. (1961-76).

☆ U.S. GOVERNMENT PRINTING OFFICE: 1980—311-493/2

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

JUNE 1980
VOL. 27, NO.
NAVSO P-1278

CONTENTS

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217